TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

─────── \* ───────

ĐỒ ÁN

**TỐT NGHIỆP ĐẠI HỌC**

NGÀNH CÔNG NGHỆ THÔNG TIN

**PHÂN VÙNG ẢNH MÀU SỬ DỤNG THUẬT TOÁN K-MEANS**

Sinh viên thực hiện: **Trần Văn Thành**

Lớp CNTT2.01 – K58

Giáo viên hướng dẫn: **PGS. Phạm Văn Hải**

HÀ NỘI 1 - 2019

# Phiếu giao nhiệm vụ đồ án tốt nghiệp

*1. Thông tin về sinh viên*

Họ và tên sinh viên: Trần Văn Thành

Điện thoại liên lạc: 0978779701

Email: thanhtranvan.hust@gmail.com

Lớp: CNTT2.01 - K58 Hệ đào tạo: Đại học chính quy

Đồ án tốt nghiệp được thực hiện tại: Viện CNTT&TT - Đại học Bách Khoa Hà Nội

Từ ngày 31/ 08/ 2018 đến ngày 28/ 12/ 2018

*2. Mục đích nội dung của ĐATN*

Phần 1: Thuật toán K-means clustering và ứng dụng cho bài toán phân vùng ảnh.

Phần 2: Phát hiện đám cháy trong ảnh và video thời gian thực.

*3. Các nhiệm vụ cụ thể của ĐATN*

1. Tìm hiểu và cài đặt các ứng dụng của K-means cho các bài toán phân loại chữ số viết tay, nén ảnh và phát hiện bất thường.
2. Cải tiến thuật toán K-means clusstering, áp dụngcho bài toán phân vùng ảnh màu, đánh giá và so sánh kết quả với các công trình nghiên cứu khác.
3. Xử lý ảnh phát hiện đám cháy trong video thời gian thực.

*4. Lời cam đoan của sinh viên*

Tôi – Trần Văn Thành - cam kết các kết quả nêu trong ĐATN là trung thực, không phải là sao chép toàn văn của bất kỳ công trình nào khác.

|  |  |
| --- | --- |
|  | *Hà Nội, ngày tháng năm 2019*  *Trần Văn Thành* |

*5. Xác nhận của giáo viên hướng dẫn về mức độ hoàn thành và cho phép bảo vệ*

|  |  |
| --- | --- |
|  | *Hà Nội, ngày tháng năm 2019*  *PSG. Phạm Văn Hải* |

# Lời cảm ơn

Lời đầu tiên, em xin chân thành cảm ơn tới thầy Phạm Văn Hải giảng viên bộ môn HTTT và thầy Đỗ Phan Thuận giảng viên bộ môn KHMT đã chấp nhận đề tài và tận tình chỉ bảo em hoàn thành đồ án này. Thầy đã cung cấp kiến thức, hướng dẫn tận tình, chỉ bảo cho em từ những bước đầu tiếp cận đề tài và trong suốt quá trình thực hiện đề tài này. Qua sự hướng dẫn của thầy em không chỉ học được nhiều kiến thức bổ ích mà còn học được những phương pháp nghiên cứu khoa học, thái độ và tinh thần nghiêm túc hết mình trong công việc.

Mặc dù em đã có những cố gắng hoàn thành đề tài này, xong do thời gian có hạn nên em cũng không thể tìm hiểu kỹ và đầy đủ hết các khía cạnh mà đề tài đề cập đến. Trong thời gian sắp tới em rất mong nhận được sự chỉ bảo tận tình của thầy hơn nữa để giúp em có thể hoàn thiện được đề tài hơn và áp dụng đề tài vào thực tế để đem lại lợi ích thiết thực theo đúng định hướng và mục tiêu mà đề tài đã đưa ra.

Em xin trân trọng cảm ơn thầy!

# CHƯƠNG I: GIỚI THIỆU

# 1.1. Tổng quan

Những năm gần đây, AI - Artificial Intelligence (Trí Tuệ Nhân Tạo), và cụ thể hơn là Machine Learning (Học Máy) nổi lên như một bằng chứng của cuộc cách mạng công nghiệp lần thứ tư 4.0 (1 - động cơ hơi nước, 2 - năng lượng điện và động cơ đốt trong, 3 – khoa học máy tính và internet). Trí Tuệ Nhân Tạo đang xuất hiện trong mọi lĩnh vực của đời sống mà có thể chúng ta không nhận ra. Xe tự lái của Google và Tesla, hệ thống tự gắng thẻ (tag) khuôn mặt người dùng của Facebook, trợ lý ảo Siri của Apple, hệ thống gợi ý sản phẩm của Amazon, hệ thống gợi ý phim của Netflix, máy chơi cờ vây AlphaGo của Google DeepMind, …, đó chỉ là một vài trong vô vàn những ứng dụng của AI/Machine Learning.

Học máy (machine learning - ML) là một tập con của AI. Nói đơn giản, Machine Learning là một lĩnh vực nhỏ của Khoa Học Máy Tính, nó có khả năng tự học hỏi dựa trên dữ liệu đưa vào mà không cần phải được lập trình cụ thể.

Sự phát triển của các hệ thống tính toán cùng với đó, lượng dữ liệu khổng lồ được thu thập bởi các hãng công nghệ lớn đã giúp Machine Learning tiến thêm một bước dài. Một lĩnh vực mới được ra đời gọi là học sâu (deep learning - DL). Deep Learning đã giúp máy tính thực thi những việc tưởng chừng như không thể vào 10 năm trước: phân loại cả ngàn vật thể khác nhau trong các bức ảnh, tự tạo chú thích cho ảnh, bắt chước giọng nói và chữ viết của con người, giao tiếp với con người, hay thậm chí cả sáng tác văn hay âm nhạc.



Mối quan hệ giữa AI, Machine Learning và Deep Learning.

Lĩnh vực machine learning và deep learning là cực kỳ rộng lớn và có nhiều nhánh nhỏ. Chúng ta luôn nhớ rằng điều đơn giản nhất là điều quan trọng nhất. Khi bắt tay vào giải quyết một bài toán machine learning hay bất cứ một bài toán nào, chúng ta nên bắt đầu từ những thuật toán đơn giản nhất. Ta không nên nghĩ rằng chỉ có những thuật toán phức tạp mới có thể giải quyết được vấn đề. Những thuật toán phức tạp thường yêu cầu độ tính toán cao và nhạy cảm với cách chọn tham số đầu vào. Thay vào đó, những thuật toán đơn giản giúp chúng ta sớm có một mô hình tổng quá cho bài toán. Các thuật toán đơn giản thường được gọi là baseline, sẽ giúp chúng ta có cái nhìn ban đầu về sự phức tạp của mỗi bài toán. Việc cải thiện kết quả sẽ được dần cải thiện ở các bước sau.

Từ những lý do trên, em quyết định sử dụng một trong những thuật toán cơ bản nhất trong Unsupervised learning - thuật toán K-means clustering (phân cụm K-means) để giải quyết bài toán.

**1.2. Mục đích đồ án**

Nghiên cứu về bài toán phân vùng ảnh (Image Segmentation) và ứng dụng của thuật toán phân cụm K-means để áp dụng nó vào giải quyết các bài toán xử lý ảnh trong thực tế. Đồng thời điều chỉnh các tham số thích hợp cho mô hình để đạt được kết quả có độ chính xác tốt nhất.

**1.3. Phát biểu bài toán**

Phần 1: Xử lý ảnh vệ tinh để tính toán được sự thay đổi của đô thị hóa, hay sự sói mòn đất, nạn chặt phá rừng, mức độ thiệt hại của thiên tai. Theo phương pháp truyền thống từ xưa dùng các bản vẽ, dựa vào bản đồ và các chuyến đi thực nghiệm, nhưng không phải nơi đâu trên trái đất con người cũng có thể đặt chân đến để kiểm tra đo đạc. Vì vậy ngày nay với các công cụ khoa học hiện đại, từ vệ tinh con người có thể thu được các hình ảnh từ bất cứ nơi đâu trên trái đất. Qua các ảnh vệ tinh chụp được một khu rừng, sau khi xử lý phân vùng ảnh thành các vùng như cây, đất, nước… Ta tính được tỷ lệ phần trăm diện tích của các vùng trong tấm ảnh, qua đó tính được diện tích các vùng, và so sánh kết quả từ đó kết luận được độ che phủ rừng, hay sự đô thị hóa… thay đổi qua các mốc thời gian.

Bài toán đặt ra: xây dựng mô hình học máy cho bài toán phân vùng ảnh (Image Segmentation) sử dụng thuật toán K-means clustering và một số ứng dụng cho các bài toán thực tế như : phân nhóm chữ số viết tay.

Phần 2: Phát hiện đám cháy trong ảnh và video thời gian thực. Hiện nay, các hệ thống báo cháy tự động trong các tòa nhà cao tầng, văn phòng,… là các cảm biến khói – thiết bị dùng để phát hiện khói – một thành phần khởi đầu của sự cháy. Có 2 loại cảm biến khói là quang điện và ion. Các cảm biến này thưởng gây ra các hiện tượng báo cháy giả và có thể làm tốn kém tiền của, khiến người dùng thiếu đi sự tin tưởng, dẫn đến hậu quả nghiêm trọng.

Bài toán đặt ra: Tách vùng cháy trong ảnh, phát hiện đám cháy trong video thời gian thực lấy từ các camera báo cháy, kết hợp với các cảm biến khói để giảm thiểu báo cháy giả.

**1.4. Dữ liệu đầu vào**

Tập dữ liệu MNIST chứa 70 000 hình ảnh về chữ số viết tay:

<http://yann.lecun.com/exdb/mnist/>

Tập dữ liệu ảnh 600 ảnh Natural Images và ảnh Ground Truth (ảnh phân đoạn bằng tay):

<https://people.eecs.berkeley.edu/~yang/software/lossy_segmentation/>

Tập dữ liệu ảnh vệ tinh thu thập từ:

<https://earthengine.google.com/timelapse/>

Tập dữ liệu ảnh có chứa đám cháy (600 ảnh) và không có đám cháy (1000 ảnh) được thu thập từ internet.

**1.5. Dữ liệu đầu ra**

Đối với bài toán phân vùng ảnh vệ tinh: ảnh đầu ra sau khi được phân vùng sẽ được gán nhãn và tỷ lệ % diện tích các vùng ảnh, đánh giá và so sánh kết quả mô hình.

Với bài toán phát hiện đám cháy trong ảnh và video realtime: đầu ra sẽ là 1 ảnh hoặc tập các ảnh là các frame được cắt từ video đầu vào sau đó được gán nhãn là cháy hoặc không cháy.

**1.6. Công cụ lựa chọn**

Thuật toán sử lý ảnh được sử dụng để giải quyết bài toán là thuật toán phân cụm K-means, được cài đặt và triển khai trên môi trường hệ điều hành Windows. Ngôn ngữ để triển khai bài toán là python v2.7 và matlab v2015b cùng thư viện sử dụng là OpenCV, và các thư viện toán học khác, các thư viện này sẽ giúp việc triển khai các hàm toán học và xử lý ảnh trong python một cách dễ dàng hơn.

**1.7. Tóm tắt nội dung đồ án**

Đồ án bao gồm phụ lục và 5 chương:

Chương 1: Giới thiệu: Chương này trình bày về nhiệm vụ cần thực hiện trong khuôn khổ đồ án tốt nghiệp, trình bày định hướng giải quyết các vấn đề đặt ra. Giới thiệu tóm tắt về cơ sở lý thuyết và công cụ được lựa chọn để giải quyết các vấn đề đặt ra.

Chương 2: Unsupervised Learning và K-means Clustering: Chương này trình bày những lý thuyết cơ bản về Unsupervised Learning , cung cấp một cách nhìn tổng quát, và những vấn đề về học không giám sát, thuật toán phân cụm K-means Clustering.

Chương 3: Image Segmentation: Chương này chủ yếu trình bày việc áp dụng thuật toán phân cụm K-means và cải tiến cho bài toán phân vùng ảnh.

Chương 4: Phát hiện đám cháy trong ảnh và video realtime: Chương này tình bày các quy tắc để phát hiện và phân loại ảnh có chứa đám cháy, áp dụng cho bài toán thực tế là phát hiện đám cháy trong video thời gian thực qua webcam trên laptop.

Chương 5: Kết luận và định hướng phát triển: Ở chương này, chúng ta đưa ra kết luận tổng kết các kết quả ban đầu, khó khăn và thuận lợi của đồ án. Các định hướng phát triển tiếp theo.

Phụ lục: Ở phần này sẽ nói về cách cài đặt, chạy chương trình giải quyết bài toán.

**Chương 2: Unsupervised Learning và K-means Clustering**

**2.1. Học không giám sát - Unsupervised Learning**

Supervised learning (học có giám sát)

Chúng ta cần nhắc lại về học có giám sát Supervised learning. Như chúng ta đã biết Supervised learning là thuật toán dự đoán đầu ra (outcome) của một dữ liệu mới (new input) dựa trên các cặp (input, outcome) đã biết từ trước. Cặp dữ liệu này còn được gọi là (data, label), tức (dữ liệu, nhãn). Supervised learning là nhóm phổ biến nhất trong các thuật toán Machine Learning. Về mặt toán học, Supervised learning là khi chúng ra có một tập hợp biến đầu vào **X =**  và một tập hợp nhãn tương ứng **Y =** , trong đó **,**  là các vector. Các cặp dữ liệu biết trước được gọi là tập *training data* (dữ liệu huấn luyện). Từ tập traing data này, chúng ta cần tạo ra một hàm số ánh xạ mỗi phần tử từ tập **X** sang một phần tử (xấp xỉ) tương ứng của tập **Y**:

**f(), i 1, 2, …, N**

Mục đích là xấp xỉ hàm số **f** thật tốt để khi có một dữ liệu **x** mới, chúng ta có thể tính được nhãn tương ứng của nó **y = f(x)**.

Thuật toán supervised learning còn được tiếp tục chia nhỏ ra thành hai loại chính:

Classification (Phân loại)

Một bài toán được gọi là *classification* nếu các *label* của *input data* được chia thành một số hữu hạn nhóm. Với **y** chỉ nhận giá trị từ một tập rời rạc, chẳng hạn {cá, cây, quả, mèo}. Ví dụ: Gmail xác định xem một email có phải là *spam* hay *non\_spam*.

Regression (Hồi quy)

Nếu *label* không được chia thành các nhóm mà là một giá trị thực cụ thể. Ví dụ: một căn nhà rộng **x** , có **y** phòng ngủ và cách trung tâm thành phố **z** *km* sẽ có giá là bao nhiêu? Với **y** nhận giá trị là số thực.

Unsupervised Learning (Học không giám sát)

Trong thuật toán này, chúng ta không biết được *outcome* hay nhãn *label* mà chỉ có dữ liệu đầu vào. Thuật toán *unsupervised learning* sẽ dựa vào cấu trúc của dữ liệu để thực hiện một công việc nào đó, ví dụ như phân nhóm (*clustering*) hoặc giảm số chiều của dữ liệu (*dimension reduction*) để thuận tiện trong việc lưu trữ và tính toán.

Về mặt toán học, *unsupervised learning* có nghĩa là khi chúng ta chỉ có dữ liệu vào **X** mà không biết nhãn **Y** tương ứng. Máy tính cần học một hàm **y = f(x)** từ tập học cho trước . **Y** có thể là các cụm dữ liệu hoặc các cấu trúc ẩn.

Những thuật toán loại này được gọi là *unsupervised learning* vì không giống như *supervised learning*, chúng ta không biết câu trả lời chính xác cho mỗi dữ liệu đầu vào.

Các bài toán Unsupervised learning được tiếp tục chia nhỏ thành ba loại:

Clustering (phân nhóm)

Một bài toán phân nhóm toàn bộ dữ liệu X thành các nhóm nhỏ dựa trên sự liên quan giữa các dữ liệu trong mỗi nhóm. Ví dụ: phân nhóm khách hàng dựa trên hành vi mua hàng. Điều này cũng giống như việc ta đưa cho một đứa trẻ rất nhiều mảnh ghép với các hình thù và màu sắc khác nhau, ví dụ tam giác, vuông, tròn với màu xanh và đỏ, sau đó yêu cầu trẻ phân chúng thành từng nhóm. Mặc dù không cho trẻ biết mảnh nào tương ứng với hình nào hoặc màu nào, nhiều khả năng chúng vẫn có thể phân loại các mảnh ghép theo màu hoặc hình dạng.

Community detection: phát hiện các cộng đồng trong mạng xã hội.

Trends detection: phát hiện xu hướng thị yếu.

**2.2. K-means Clustering**

**2.2.1. Giới thiệu**

Trong thuật toán K-means clustering, chúng ta không biết nhãn *label* của từng điểm dữ liệu. Mục đích là làm thế nào để phân dữ liệu thành các cụm *cluster* khác nhau sao cho dữ liệu trong cùng một cụm có tính chất giống nhau.

Ý tưởng đơn giản nhất về *cluster* là tập hợp các điểm ở gần nhau trong một không gian nào đó (không gian này có thể có rất nhiều chiều trong trường hợp thông tin về một điểm dữ liệu là rất lớn). Hình bên dưới là một ví dụ về 3 cụm dữ liệu trong không gian 2 chiều.



**2.2.2. Phân tích toán học**

Mục đích cuối cùng của thuật toán phân nhóm này là: từ dữ liệu đầu vào và số lượng nhóm chúng ta muốn tìm, hãy chỉ ra *center* của mỗi nhóm và phân các điểm dữ liệu vào các nhóm tương ứng. Giả sử rằng mỗi điểm dữ liệu chỉ thuộc vào đúng một nhóm.

Giả sử có **N** điểm dữ liệu là **X =**  và **K N** là số cluster chúng ta muốn phân chia. Chúng ta cần tìm các center **, , …,** và label của mỗi điểm dữ liệu.

Với mỗi điểm dữ liệu đặt **=** là label vector của nó, trong đó nếu được phân vào *cluster* **k** thì **= 1** và **= 0**, . Điều này có nghĩa là có đúng một phần tử của vector là bằng 1 (tương ứng với cluster của ), các phần tử còn lại bằng 0. Ví dụ: nếu một điểm dữ liệu có *label* vector là thì nó thuộc vào *cluster* 1, nếu là thì nó thuộc vào *cluster* 2, .. Ràng buộc của có thể viết dưới dạng toán học như sau:

Hàm mất mát và bài toán tối ưu

Nếu ta coi center là center (hoặc representative) của mỗi *cluster* và ước lượng tất cả các điểm được phân vào *cluster* này bởi , thì một điểm dữ liệu được phân vào cluster ***k*** sẽ bị sai số là . Ta mong muốn sai số này có trị tuyệt đối nhỏ nhất nên sẽ tìm cách để đại lượng sau đây đạt giá trị nhỏ nhất:

Hơn nữa, vì được phân vào *cluster* ***k*** nên **= 1, = 0,** . Khi đó, biểu thức bên trên sẽ được viết lại là:

Sai số cho toàn bộ dữ liệu sẽ là:

Trong đó **Y =** , **M =** lần lượt là các ma trận được tạo bởi *label* vector của mỗi điểm dữ liệu và *center* của mỗi *cluster*. Hàm số mất mát trong bài toán K-means clustering là hàm với ràng buộc như được nêu trong phương trình .

Ta cần tối ưu bài toán sau:

Ta cần tối ưu 2 biến số Y và M để hàm mất mát *đạt giá trị nhỏ nhất*.

Thuật toán tối ưu hàm mất mát

Bài toán là một bài toán khó tìm điểm tối ưu vì nó có thêm các điều kiện ràng buộc. Bài toán này thuộc loại *mix-integer programming* (điều kiện biến là số nguyên) - là loại rất khó tìm nghiệm tối ưu toàn cục (*global optimal point*, tức nghiệm làm cho hàm mất mát *đạt giá trị nhỏ nhất* có thể). Tuy nhiên, trong một số trường hợp ta vẫn có thể tìm được phương pháp để tìm được nghiệm gần đúng hoặc điểm cực tiểu. Điểm cực tiểu ở đây chưa chắc đã phải là điểm làm cho hàm số đạt giá trị nhỏ nhất (cực tiểu toàn cục), mà nó có thể là điểm cực tiểu địa phương (điểm cực tiểu cục bộ).

Một cách đơn giản để giải bài toán là xen kẽ giải Y và M khi biến còn lại được cố định làm tham số.

Cố định M và tìm Y

Giả sử đã tìm được các centers, hãy tìm các label vector để hàm mất mát đạt giá trị nhỏ nhất. Điều này tương đương với việc tìm cluster cho mỗi điểm dữ liệu.

Khi các centers là cố định, bài toán tìm label vector cho toàn bộ dữ liệu có thể được chia nhỏ thành bài toán tìm label vector cho từng điểm dữ liệu như sau:

Vì chỉ có một phần tử của *label* *vector* bằng 1 nên bài toán có thể tiếp tục được viết dưới dạng đơn giản hơn:

Vì chính là bình phương khoảng cách tính từ điểm đến *center* , ta có thể kết luận rằng mỗi điểm thuộc vào *cluster* có *center* gần nó nhất! Từ đó ta có thể dễ dàng suy ra *label vector* của từng điểm dữ liệu.

Cố định Y và tìm M

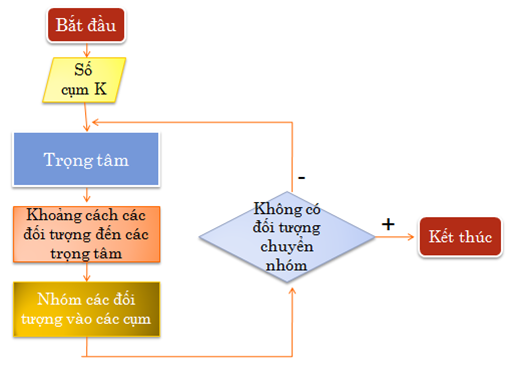
Giả sử đã tìm được cluster cho từng điểm, hãy tìm center mới cho mỗi cluster để hàm mất mát đạt giá trị nhỏ nhất. Một khi chúng ta đã xác định được label vector cho từng điểm dữ liệu, bài toán tìm center cho mỗi cluster được rút gọn thành:

Lúc này, ta có thể tìm nghiệm bằng phương pháp giải phương trình đạo hàm bằng 0, vì hàm cần tối ưu là một hàm liên tục và có đạo hàm xác định tại mọi điểm. Và quan trọng hơn, hàm này là hàm lồi (convex) theo nên chúng ta sẽ tìm được giá trị nhỏ nhất và điểm tối ưu tương ứng.

Giải phương trình đạo hàm bằng 0 ta có:

Ở đây mẫu số chính là phép đếm số lượng các điểm dữ liệu trong cluster ***j***. Còn tử số chính là tổng các điểm dữ liệu trong cluster ***j***. Nghĩa là chính là trung bình cộng của các điểm trong cluster ***j***.

**2.2.3. Thuật toán K-means**



Đầu vào: Dữ liệu X và số cụm cluster cần tìm K.

Đầu ra: Các center M và label vector cho từng điểm dữ liệu Y.

1. Chọn K điểm bất kỳ làm các center ban đầu.
2. Phân mỗi điểm dữ liệu vào cluster có center gần nó nhất.
3. Nếu việc gán dữ liệu vào từng cluster ở bước 2 không thay đổi so với vòng lặp trước nó thì ta dừng thuật toán.
4. Cập nhật center cho từng cluster bằng cách lấy trung bình cộng của tất các các điểm dữ liệu đã được gán vào cluster đó sau bước 2.
5. Quay lại bước 2.

Thuật toán sẽ dừng lại sau một số hữu hạn vòng lặp (luôn hội tụ sau một số hữu hạn lần lặp). Vì hàm mất mát là một số dương và sau mỗi bước 2 hoặc 3, giá trị của hàm mất mát bị giảm đi. Số lượng cách phân nhóm cho toàn bộ dữ liệu là hữu hạn nên đến một lúc nào đó, hàm mất mát sẽ không thể thay đổi, và khi đó thuật toán sẽ dừng.

**Chương 3: Image Segmentation**

Trước khi đi vào bài toán chính, tôi sẽ đi vào một bài toán cũng rất thực tế và cơ bản sau, để có thể thấy rõ hơn cách biểu diễn dữ liệu và xây dựng mô hình cho bào troán phân vùng ảnh với thuật toán phân cụm K-means clusstering.

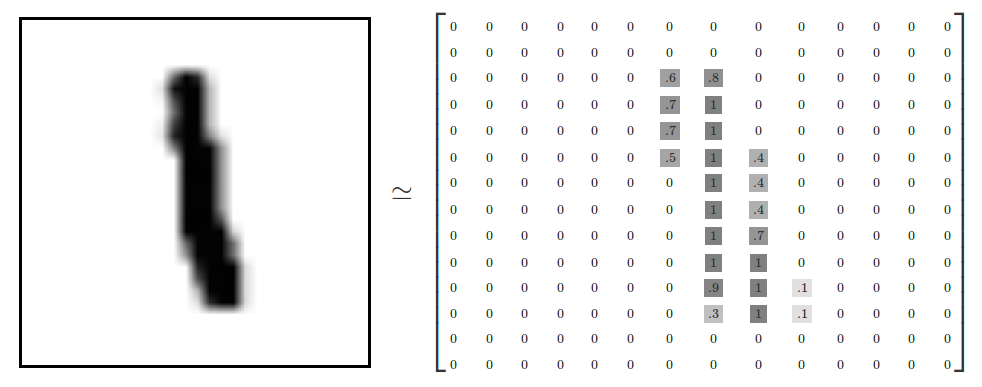
**3.1. Phân nhóm chữ số viết tay**

**3.1.1. Biểu diễn dữ liệu**

Trước hết ta đi vào bài toán Image Classification cho bộ cơ sở dữ liệu chữ số viết tay MNIST gồm hai tập con: tập dữ liệu huấn luyện (*training\_set*) có tổng cộng 60000, và tập dữ liệu thử nghiệm (*test\_set*) có tổng cộng 10000 ví dụ khác nhau. Tất cả đều đã được gán nhãn.

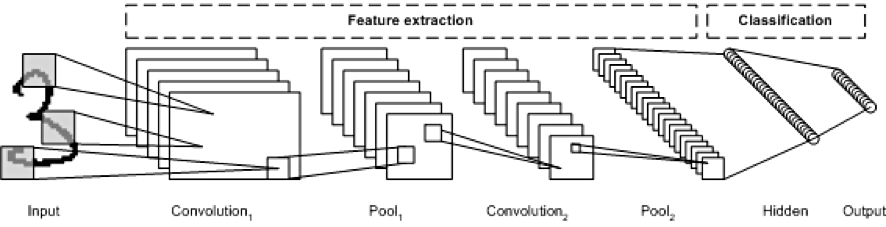


Mỗi bức ảnh là một ảnh đen trắng (có 1 channel), có kích thước 28x28 pixel. Khi đó mỗi bức ảnh sẽ là 1 vector 1x784, mà mỗi phần tử của vector là một pixel mang một giá trị là một số tự nhiên từ 0 đến 255. Các pixel màu đen có giá trị bằng 0, các pixel càng trắng thì có giá trị càng cao. Tất cả các chữ số được đặt trên nền đen, và chữ số màu sáng. Dưới đây là một ví dụ về chữ số 7 và giá trị các pixel của nó (ảnh nền đã được thay đổi thành nền trắng).



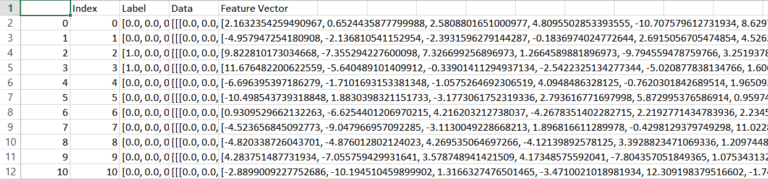
Ta coi mỗi bức ảnh là một điểm dữ liệu. Và mỗi điểm dữ liệu là 1 vector có kích thước 1x784 biểu diễn 1 bức ảnh chữ số.

Ở đây chúng ta sẽ sử dụng các vector đặc trưng được trích xuất bởi các mô hình CNN để phân cụm dữ liệu MNIST:



Sơ đồ kiến trúc mạng LeNet

Trích xuất đặc trưng từ ảnh thu được các vector đặc trưng và lưu vào file .csv có dạng như sau:



**3.1.2. K-Means clustering of MNIST data**

Dự đoán số cụm K

Trước khi áp dụng thuật toán K-means clustring cho bộ dữ liệu ta cần dự đoán số cụm K thích hợp.

Phương pháp khuỷu tay (Elbow): ý tưởng là đối với một phạm vi các giá trị thích hợp của K, ta tính sai số của tổng lỗi trong cụm (wcss - within-cluster sum of errors) để tìm ra giá trị K tối ưu nhất. Ta có thể tìm thấy K ở vị trí khuỷu tay, tức là trước khuỷu tay lỗi giảm nhanh dần và sau khi qua khuỷu tay lỗi giảm không đáng kể.

Phương pháp Average Silhouette: phương pháp này thể hiện tính nhất quán trong các *cluster*, giá trị *silhouette* là thước đo mức độ tương tự của một đối tượng với cụm (sự gắn kết) của chính nó so với các cụm khác (sự tách biệt) hay mức độ phân tách giữa các cụm. Silhouette nằm trong phạm vi −1 đến +1, trong đó giá trị cao biểu thị rằng đối tượng được kết hợp tốt với cụm của chính nó và tách rời với các cụm lân cận. Nếu hầu hết các đối tượng có giá trị cao, thì cấu hình phân cụm là phù hợp. Nếu nhiều điểm có giá trị thấp hoặc âm, thì cấu hình phân cụm có thể có quá nhiều hoặc quá ít cụm. Silhouette được tính như sau:

Tính khoảng cách trung bình từ tất cả các điểm dữ liệu trong cùng một cụm ().

Tính khoảng cách trung bình từ tất cả các điểm dữ liệu trong cụm gần nhất ().

Tính hệ số:

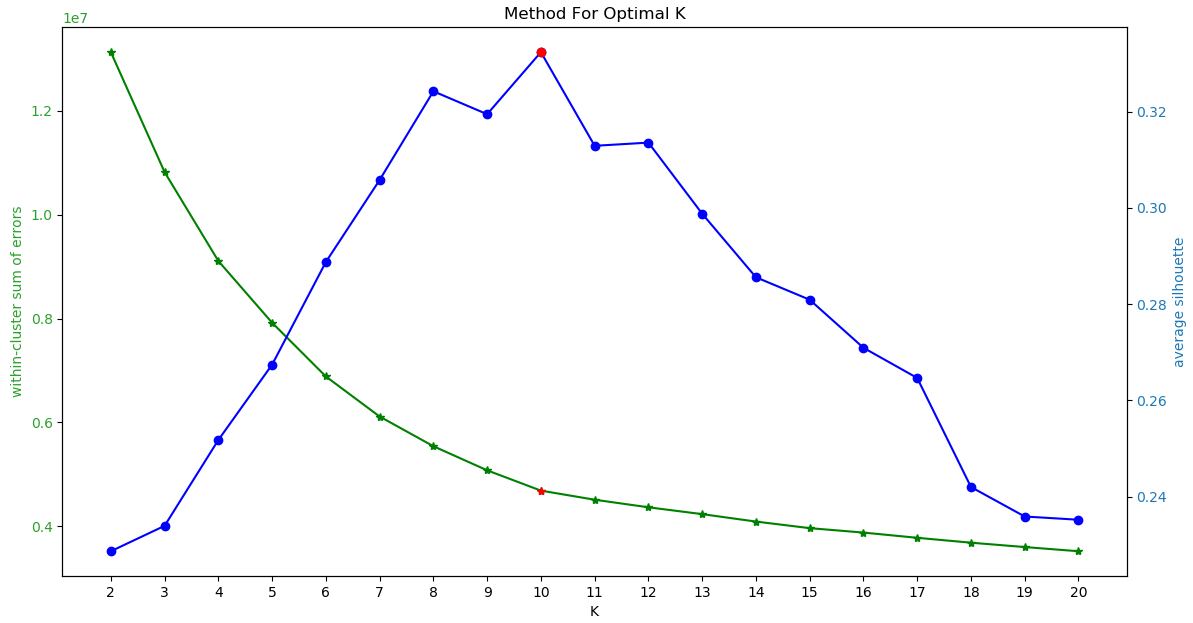
Hoặc là

Nếu mẫu i rất gần với các cụm lân cận.

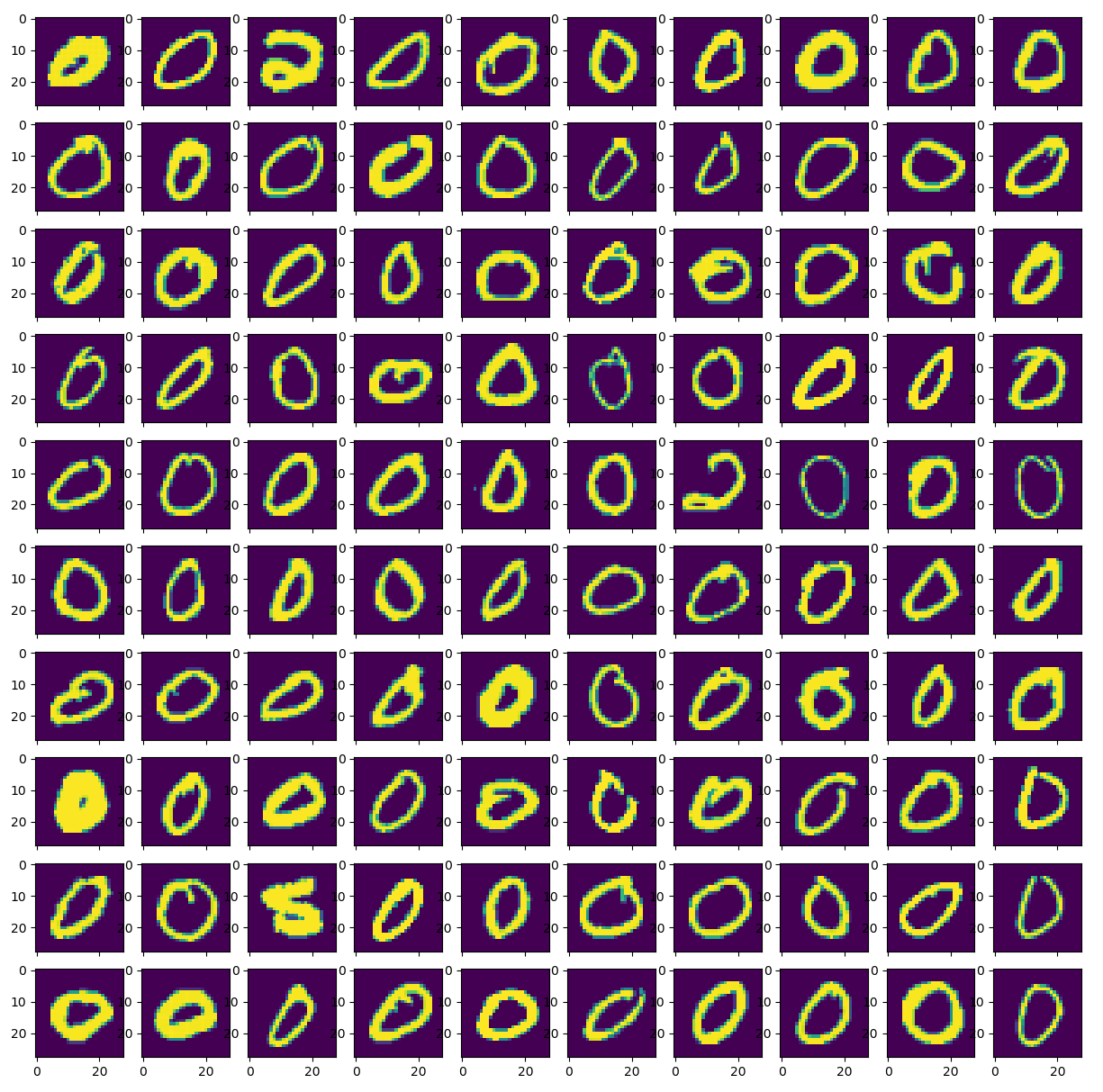
Nếu mẫu nằm cách xa các cụm lân cận.

Nếu mẫu được gán cho các cụm sai.

Dưới đây là kết quả khi thực hiện tối ưu số cụm K, sử dụng 2 phương pháp trên, kết quả khá rõ khi chọn số cụm K = 10 là tốt nhất.



Triển khai thuật toán K-means clustering cho bộ dữ liệu features matrix ta được kết quả:



Đánh giá độ chính xác

Trong đó: là số chữ số giống nhau nhiều nhất trong cụm *j*

K là số cụm, N tổng số bức ảnh, là số bức ảnh thuộc cụm *j*.

Tức là ta chia số lượng những chữ số được phân đúng (số lượng nhiều nhất trong 1 cụm) chia cho tổng các chữ số trong cụm đó. Trong hình trên, ta hiển thị 1 cụm có kích thước 100 chữ số, gồm có 97 chữ số 0 và 3 chữ số bị phân nhóm sai là 2 chữ số 2 và 1 chữ số 5. Như vậy độ chính xác của cụm này là 97%. Dưới đây là kết quả phân loại trên toàn bộ tập dữ liệu:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cluster | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Accuracy (%) | 94.68 | 93.54 | 80.69 | 91.56 | 96.71 | 92.66 | 90.10 | 86.82 | 94.84 | 85.47 |
| Average (%) | 90.71 | | | | | | | | | |
| Time  (s) | 4.640 | | | | | | | | | |

Nhận thấy kết quả ở mức khá tốt khi thực hiện phân loại chữ số viết tay bằng K-meas clustering, so với các kết quả từ các bài báo khác sử dụng mô hình CNN sẽ có độ chính xác cao từ 95-98% với cùng 1 tập dữ liệu MNIST.

**3.2. Image Segmentation và cải tiến K-means**

**3.2.1. Bài toán phân vùng ảnh (image segmentation)**

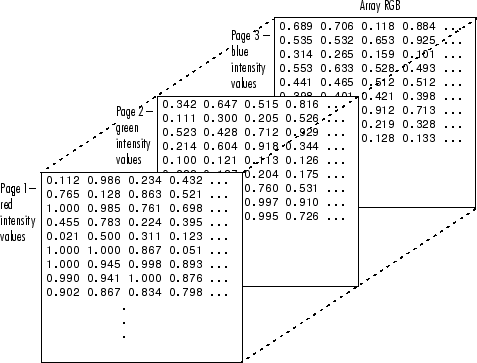
Thuật toán K-means đã được sử dụng rất nhiều trong nghiên cứu và được cài đặt trong các phần mềm xử lý ảnh viễn thám. Tuy nhiên, khi phân vùng ảnh viễn thám kích thước lớn, tốc độ hội tụ của thuật toán vẫn rất chậm. Trong đồ án này, tôi đề xuất một cải tiến thuật toán phân cụm K-Means cải tiến cho việc dữ đoán và khởi tạo tâm hiệu quả, kết hợp với bước tiền xử lý ảnh đầu vào nhằm tăng tốc độ phân vùng ảnh viễn thám có kích thước lớn, và cho kết quả đầu ra tốt đối với ảnh bị nhiễu.

**3.2.2. Biểu diễn dữ liệu**

Ý tưởng cơ bản để phân đoạn hình ảnh bằng thuật toán phân cụm K-means là tạo các cụm dựa trên giá trị màu của mỗi pixel, vì thế độ chói (hay độ sáng) của mỗi pixel sẽ bị bỏ qua, ta chỉ quan tâm đến các thành phần về mầu của pixel. Chính vì thế nên trong phần này, ta áp dụng phân cụm K-means cho ảnh trên không gian mầu L\*a\*b\*. Trong không gian màu L\*a\*b\* ta loại bỏ thành phần về độ sáng L, và chỉ giữ lại 2 thành phần về mầu là kênh mầu a và b. Lúc này kích thước dữ liệu đầu vào giảm, giúp cho thuật toán chạy nhanh hơn và phân đoạn ảnh chính xác hơn. Khi đó ta cẩn *reshape* lại kích thước của ma trận ảnh từ ma trận 3 chiều sang ma trận 2 chiều. Ảnh số thực chất là một ma trận các điểm ảnh. Đối với ảnh màu, mỗi điểm ảnh sẽ có 3 thuộc tính tương ứng với 3 chiều dữ liệu. Hai chiều về không gian chính là tọa độ hay vị trí của pixel đó trong ma trận điểm ảnh, chiều thứ ba là giá trị về màu sắc (màu ở đây có thể là không gian màu RGB) của pixel.

Trong đó: *h, w* là kích thước và *c* là số kênh màu của ảnh *I*.

Mỗi điểm ảnh lúc này sẽ là 1 hàng của ma trận A.



Công thức chuyển đổi từ hệ màu RGB sang hệ màu L\*a\*b\*:

Trong đó:

và

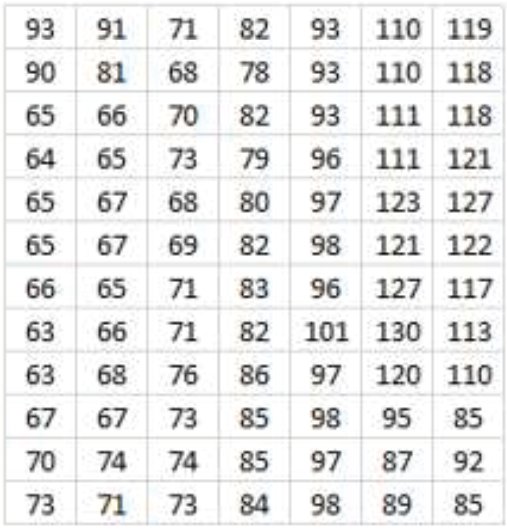
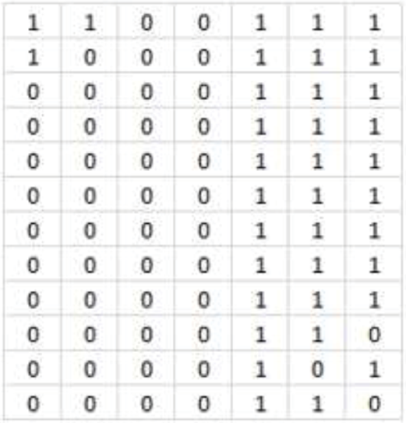
**3.2.3. Tiền xử lý**

Là quá trình sử dụng các kỹ thuật xử lý ảnh để nâng cao chất lượng ảnh. Phương pháp sử dụng SMQT (Successive Mean Quantization Transform), nhằm xóa bỏ sự chênh lệch giữa các cảm biến do độ tăng và độ sai lệch. SMQT có thể được sử dụng để mở rộng biểu diễn cấu trúc đến số bit được xác định trước tùy ý trên dữ liệu chiều tùy ý.

Đơn vị cơ bản của SMQT là MQU (Mean Quantization unit) trong đó bao gồm việc tính giá trị trung bình của tất cả các pixel trong ảnh, sau đó giá trị trung bình được sử dụng để định lượng giá trị của dữ liệu thành 0 hoặc 1, tùy thuộc vào nếu giá trị của pixel thấp hơn hoặc cao hơn giá trị trung bình.

MQU:

Pixel (V) Mean

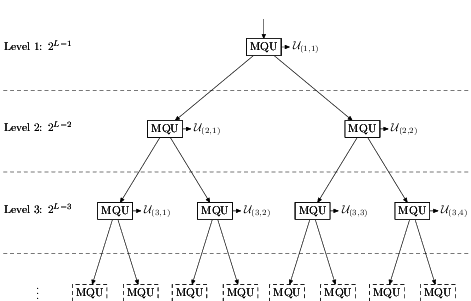
 

Mean = 87.5595

SMQT có thể được coi là một cây của các phép tính MQU, trong đó trọng số

được đưa ra tùy thuộc vào mức độ hiện tại của cây.

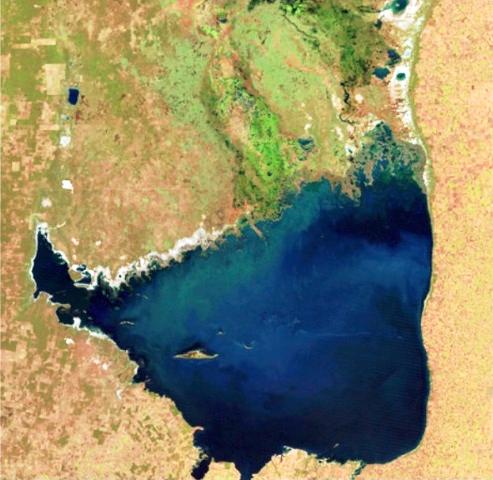
Trong đó L là mức tối đa của cây, i là mức hiện tại.



Trong ảnh RGB, SMQT có thể được áp dụng theo hai cách khác nhau. Đầu tiên, áp dụng SMQT cho ba kênh RGB. Cho là tất cả các giá trị dữ liệu bất kể kênh nào, thì:

Điều này sẽ dẫn đến sự tăng cường độ tương phản phi tuyến bảo tồn thứ tự của các giá trị RGB cho mỗi pixel, nhưng với khoảng cách thay đổi giữa các giá trị R, G và B trong mỗi pixel. Cách thứ hai là áp dụng nó một cách riêng biệt trong mỗi ba kênh. Đặt , và là tất cả các giá trị dữ liệu của kênh màu R, G và B tương ứng, thì:

Cuối cùng, các pixel được tăng cường thu được bằng cách ghép các kênh màu , và . Dưới đây là hình ảnh sau quá trình nâng cao chất lượng ảnh bằng phương pháp SMQT:

**3.2.4. Xây dựng mô hình**

Hiệu năng của thuật toán K-means bị ảnh hưởng rất nhiều ở bước khởi tạo trung tậm K cụm đầu tiên. Tùy vào các center ban đầu mà thuật toán có thể có tốc độ hội tụ rất chậm, hoặc thậm chí cho chúng ta nghiệm không chính xác (chỉ là local minimum - điểm cực tiểu địa phương - mà không phải giá trị nhỏ nhất). Đó là một nhược điểm lớn của K-means cho bài toán phân đoạn ảnh nói riêng và phân cụm nói chung. Vì thế ta sẽ đi khắc phục nhược điểm này để có được khởi tạo trung tâm K ban đầu tốt nhất.

Một nhược điểm nữa ta sẽ cần khắc phục đó là K-means cho kết quả không tốt đối với các hình ảnh bị nhiễu, hay có chất lượng thấp. Vì thế trước khi thực hiện phân cụm ảnh với K-means, ta sẽ nâng cao chất lượng ảnh đầu vào, tăng cường đặc trưng cho các điểm ảnh.

Thuật toán K-means được chứng minh là hội tụ và có độ phức tạp tính toán là O(tkn)  
với **t** là số lần lặp, **k** là số cụm, **n** là số đối tượng của tập dữ liệu vào. Thông thường  
**k<<n** và **t<<n** thường kết thúc tại một điểm tối ưu cục bộ.

Công việc cải tiến phương pháp cho bài toán phân vùng ảnh màu sử dụng K-means clusstering:

1. Đọc ảnh màu RGB.
2. Nâng cao chất lượng ảnh dừng phương pháp SMQT.
3. Chuyển đổi hình ảnh từ định dạng RGB sang Lab.
4. Định hình lại ảnh thành ma trận hai chiều. Vì ảnh ban đầu ở không gian 3D, ta chuyển về không gian 2D (Trong không gian L\*a\*b loại bỏ kênh độ sáng L, chỉ lấy 2 kênh màu a\* và b\*).
5. Áp dụng phương pháp Elbow và Silhouette để dự đoán số cụm K từ ma trận dữ liệu.
6. Sắp xếp ma trận dữ liệu ma trận hình ảnh theo thứ tự không giảm, chia toàn bộ dữ liệu hình ảnh được sắp xếp thành K phần (K đã được dự đoán ở bước 4).
7. Khởi tạo K trung tậm cụm đầu tiên: Tìm giá trị trung bình của từng phần và giá trị trung bình đó trở thành trọng tâm cho cụm tương ứng và lưu lại trung tậm cụm ban đầu.
8. Áp dụng thuật toán phân cụm K-means cho ma trận dữ liệu hình ảnh với trung tâm cụm ban đầu và số cụm K đã dự đoán ở trên.

**3.2.5. Kết quả thử nghiệm và đánh giá mô hình**

Các thuật toán phân cụm có thể khó đánh giá mà không có trước thông tin hoặc giả định về dữ liệu. Vì vậy trước khi thực hiện thuật toán phân cụm, ta sẽ phân tích và dự đoán số cụm K từ dữ liệu ảnh đầu vào.

Kết quả chạy thử và so sánh với hình ảnh truth image:

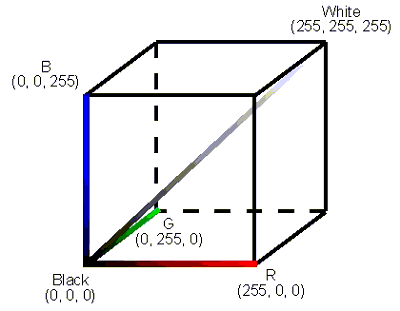
Kết quả chạy với bộ dữ liệu ảnh vệ tinh:

**Chương 4: Phát hiện đám cháy trong Ảnh và Video**

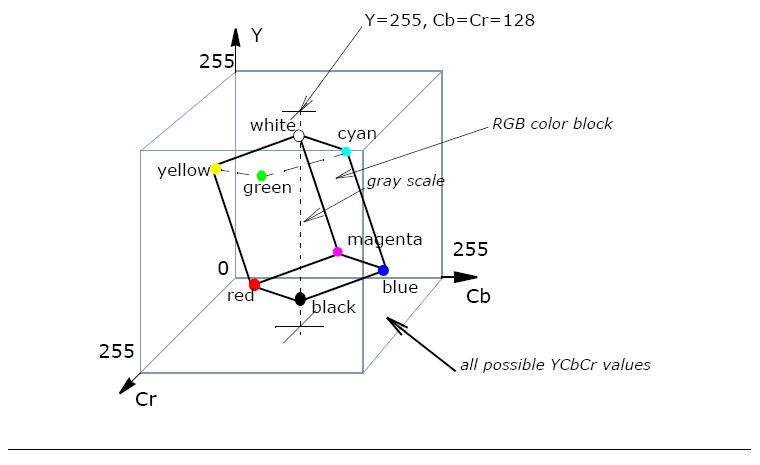
**4.1. Biểu diễn dữ liệu**

Sử dụng không gian màu RGB và YCbCr, ưu điểm khi sử dụng không gian màu YCbCr là nó có thể tách độ sáng hiệu quả hơn không gian màu RGB.

Một ảnh số trong không gian màu RGB có 3 mặt phẳng R, G, B. Các sự kết hợp giữa các mặt phẳng màu RGB cho phép các thiết bị hiển thị biểu diễn màu trong môi trường kỹ thuật số. Mỗi mặt phẳng màu được lượng hóa thành các mức rời rạc. Nói chung có 256 (8 bit cho mỗi mặt phẳng màu) lượng tử hóa được sử dụng cho mỗi mặt phẳng, ví dụ màu trắng được biểu thị bằng (R, G, B) = (255,255,255) và màu đen được biểu thị bằng (R, G, B) = (0, 0, 0). Một ảnh màu bao gồm nhiều pixel, trong đó mỗi pixel được biểu thị bằng vị trí không gian trong lưới hình chữ nhật (x, y) và vector màu (R (x, y), G (x, y), B (x, y)) tương ứng với vị trí không gian (x, y) là vị trí tọa độ của pixel.



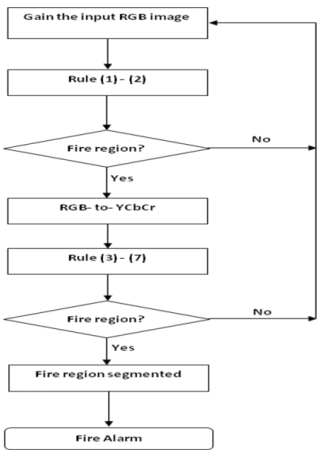
Công thức chuyển đổi từ hệ màu RGB sang hệ màu YCbCr:



Tập dữ liệu kiểm thử gồm 2 tập con, tập *test\_fire* chứa 600 ảnh có đám cháy và tập *non\_fire* chứa 1000 ảnh không có đám cháy, tất cả đều đã được gán nhãn.

**4.2. Xây dựng mô hình**

Sơ đồ thuật toán:



*Nếu 1 điểm là điểm ảnh cháy khi nó thỏa mãn đồng thời 7 quy tắc sau đây:*

**Quy tắc 1**

Có thể nhận thấy các vùng cháy trong ảnh, kênh R có giá trị cường độ cao hơn kênh G và kênh G có giá trị cường độ cao hơn giá trị cường độ kênh B.

Hình ảnh được phân vùng các điểm chớp cháy của nó như hình 1 với kênh màu green. Sau đó chúng ta tính giá trị trung bình của các mặt phẳng R, G và B trong các vùng lửa được phân vùng của ảnh gốc. Vì vậy, đối với pixel tại vị trí không gian (x, y) sẽ là pixel lửa nếu quy tắc bên dưới phải được thỏa mãn với pixel này.

**Quy tắc 2**

Xác định một số giá trị ngưỡng cho điểm ảnh lửa:

Khi hình ảnh được chuyển đổi từ không gian màu RGB sang YCbCr, cường độ và không gian màu sẽ dễ ràng phân biệt. Điều này giúp việc nhận dạng vùng lửa dễ dàng trong không gian màu YCbCr.

Trong đó Y là độ sáng, Cb và Cr là các thành phần độ màu Blue và độ màu Red tương ứng. Với ảnh màu RGB, nó được chuyển thành hình ảnh màu YCbCr bằng cách sử dụng RGB chuyển YCbCr.

Giá trị trung bình của ba thành phần Y, Cb và Cr, được biểu thị bằng Ymean, Cbmean và Crmean tương ứng được tính như sau:

Trong đó, (x, y) là vị trí tọa độ của pixel trong ảnh, M × N

là tổng số pixel trong hình ảnh đã cho

**Quy tắc 3 và 4**

**Quy tắc 5**

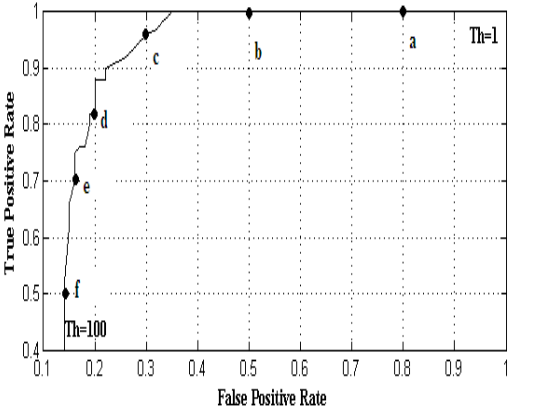
Vùng ngọn lửa nói chung là vùng sáng nhất trong bối cảnh nhìn thấy của bức ảnh hay camera. Giá trị trung bình trên toàn bộ tấm ảnh của ba kênh Ymean, Cbmean và Crmean chứa thông tin có giá trị. Đối với vùng ngọn lửa, giá trị của thành phần Y là lớn hơn thành phần Y trung bình trên toàn bộ bức ảnh. Trong khi giá trị của thành phần Cb tại vùng lửa nhỏ hơn giá trị Cb trung bình của trên toàn bộ bức ảnh. Hơn nữa, thành phần Cr của vùng lửa lớn hơn Cr trung bình trên toàn bộ bức ảnh:

Trong đó, chỉ ra rằng bất kỳ điểm ảnh nào thỏa mãn điều kiện được đưa ra trong phương trình (5) được coi là điểm ảnh cháy.

**Quy tắc 6**

Có sự khác biệt đáng kể giữa Cb và các thành phần Cr của các điểm ảnh cháy. Đối với pixel lửa, thành phần Cb chủ yếu là "đen" (cường độ thấp hơn) trong khi thành phần Cr chủ yếu là "trắng" (cường độ cao hơn). Thực tế này có thể được dịch sang một quy tắc khác như sau:

Giá trị của Th được chọn sao cho tỷ lệ phát hiện vượt quá 95% và tỷ lệ báo động giả nhỏ hơn 30% (điểm c) tương ứng với Th = 70.



**Quy tắc 7**

Chúng ta có đã xác định một số giá trị ngưỡng cho điểm ảnh lửa. Ta chỉ xét giá trị ngưỡng cho 2 mặt phẳng Cb và Cr, không xét mặt phẳng Y vì nó là thành phần độ sáng và nó phụ thuộc vào điều kiện chiếu sáng.

Một điểm ảnh là pixel vùng lửa nếu thỏa mãn đồng thời 7 quy tắc trên.

|  |  |
| --- | --- |
| Color space | Rules |
| RGB | 1. RGB |
| 1. R Rmean G Gmean B Bmean |
| YCbCr | 1. Y(x,y) Cb(x,y) |
| 1. Cr(x,y) Cb(x,y) |
| 1. Y(x,y) Ymean Cb(x,y) Cbmean Cr(x,y) Crmean |
|  |
| 1. (Cb(x,y) 120) (Cr(x,y) 150) |

Kết quả khi thử nghiệm 1 số ảnh:

**Phát hiện vùng khói**

Tương tự như phát hiện vùng lửa, ta có thể mô hình hóa các điểm ảnh khói. Nhưng điểm ảnh khói không hiển thị các đặc điểm màu sắc như pixel lửa. Lúc đầu, khi nhiệt độ khói thấp, dự kiến khói sẽ hiển thị từ màu trắng-xanh đến trắng. Về sau nhiệt độ của khói tăng lên và màu chuyển từ màu xám đen đến màu đen. Vì vậy, tôi có thể xây dựng các điểm ảnh khói như sau:

*Th*

*Th*

*Th*

Trong đó Th là ngưỡng toàn cầu dao động từ 15 đến 25, các điểm ảnh khói nên có tương tự cường độ trong các kênh màu RGB của chúng. Vì thông tin khói sẽ được sử dụng cho hệ thống phát hiện cháy sớm nhất, nên phát hiện các mẫu khói khi khói có nhiệt độ thấp. Đây là trường hợp, mà các mẫu khói có màu từ trắng-xanh đến trắng, có nghĩa là độ bão hòa của màu sắc nên càng thấp càng tốt. Sử dụng ý tưởng này, quy tắc sau được sử dụng khi không gian màu HSV được sử dụng:

S(x, y) 0.1

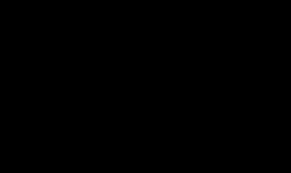
Để phân biệt được khói và các hình ảnh tương tự như sương mù hay mây trời, ta sẽ phát hiện sự chuyển động của khói và lửa trong video đám cháy.

**4.3. Kết quả thử nghiệm và đánh giá**

Kết quả khi chạy thử trên bộ dữ liệu thử nghiệm gồm 1000 ảnh chứa đám cháy và 5000 ảnh không chứa đám cháy, tất cả đều đã được gán nhãn.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |

Dưới đây là hình ảnh khi phát hiện có đám cháy, và không phát hiện ra đám cháy đối với những hình ảnh không chứa ngọn lửa nhưng có màu sắc của ngọn lửa như mặt trời, ánh đèn sáng chói,…

Nội dung đồ án bao gồm các phần chính:

# Kết quả thử nghiệm

## Mô tả dữ liệu

Trong các thí nghiệm của chúng tôi, chúng tôi sử dụng các tập dữ liệu được sử dụng trong [17] [10], được lấy từ miRbase [46], NCBI (<https://www.ncbi.nlm.nih.gov/>), fRNAdb [33], NONCODE [47] và snoRNA-LBME-db [48]. Các bộ dữ liệu được gọi là *human, cross-spieces, new.*

Bảng 1 cho thấy thống kê số lượng mẫu “positive” và “negative” trong mỗi bộ dữ liệu. Tổng cộng, có 3230 mẫu “positive” và 23934 mẫu “negative”, dẫn đến sự mất cân bằng trong các lớp. Để đối phó với sự mất cân bằng của lớp, chúng tôi gán trọng số nhiều hơn cho các mẫu “positive” trong hàm mất mát cross entropy. Hình 10 cho thấy biểu đồ độ dài chuỗi trong các tập dữ liệu. Trong khi chiều dài trung bình của chuỗi RNA là 96, thì vẫn có chuỗi với hơn 300 nucleotide. Trình tự RNA dài nhất và ngắn nhất tương ứng là 398 và 45 nucleotide.

Trong giai đoạn xây dựng mô hình, 90% bộ dữ liệu *human* và *cross-spieces* được sử dụng. Chúng tôi chọn kiến ​​trúc của mô hình và hyperparameters (tức là số epochs, learning rate, batch size) dựa trên kết quả của 5-fold cross validation (cross validation được chạy trên 90% của tập *human* và trên 90% của tập *cross-spieces*, nhưng chúng tôi đã tìm thấy các hyperparameters nói trên hoạt động tốt cho cả hai), đó là quy trình chuẩn để điều chỉnh hyperparameters trong tài liệu nhận dạng miRNA trước [46] [10].

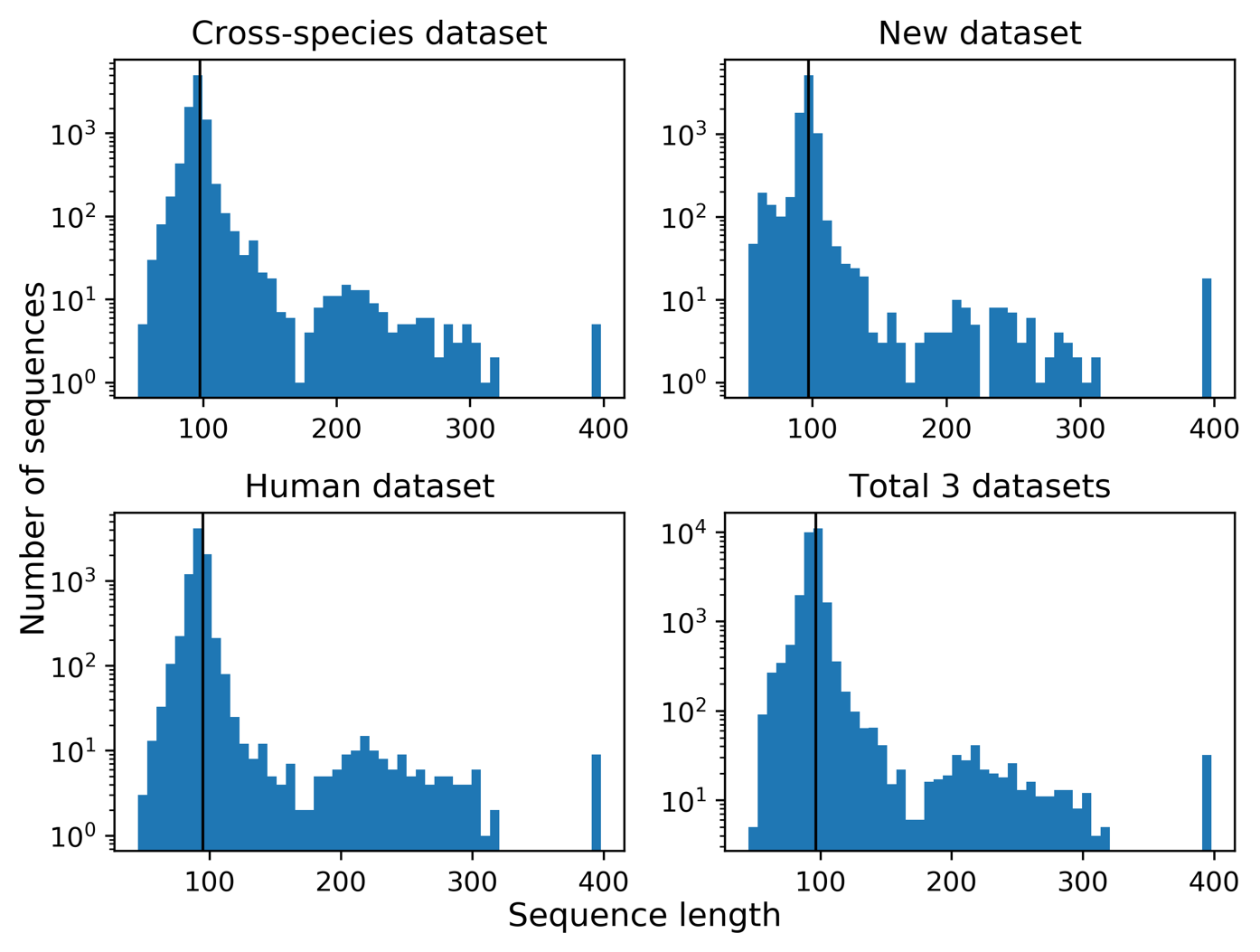
Sau khi thực hiện xong bước cross validation, chúng tôi sử dụng ba thủ tục đánh giá (testing) tiêu chuẩn:

1. Train trên 90% trên bộ *human*, test 10% còn lại của bộ dữ liệu *human*, như được thực hiện trong [17].
2. Train trên 90% trên bộ *cross-spieces*, test 10% còn lại của tập *cross-spieces*, như đã hoàn thành trong [17].
3. Train trên toàn bộ dữ liệu cross-spieces, test toàn bộ tập dữ liệu *new*, như đã đề xuất.

Chúng tôi đã không sử dụng thử nghiệm 10% bộ dữ liệu *human* và *cross-spieces*, cũng như bất kỳ mẫu nào từ tập dữ liệu *new* tại bất kỳ thời điểm nào trong quá trình training cũng như tìm kiếm hyperparameters.

|  |  |  |  |
| --- | --- | --- | --- |
|  | human | Cross-species | new |
| Positive samples | 863 | 1677 | 690 |
| Negative samples | 7422 | 8266 | 8246 |

Bảng 1. Thống kê của các tập dữ liệu



Hình 10. Thống kê dữ liệu

Chúng tôi đánh giá và so sánh kết quả của mô hình với các nghiên cứu trước đó trong lĩnh vực này. Như đã mô tả ở trên, chúng tôi đã thử nghiệm trên 10% bộ dữ liệu *human, cross-spieces*, và tất cả các tập dữ liệu *new*. Chúng tôi cũng báo cáo kết quả 5-fold cross validation, cũng là một đánh giá phổ biến trong học máy. Để so sánh, chúng tôi sử dụng các chỉ số sensitivity (SE), specificity (SP), positive predictive value (PPV), F-score, g-mean, area under the receiver operating characteristic (AUROC), and area under the precision-recall curve (AUPR). Các kết quả được thể hiện trong Bảng 2 Bảng 3 Bảng 4. Các kết quả của miRBoost, CSHMM, triplet-SVM, microPred, MIReNA và deepMiRGene được lấy từ [17]. Chúng tôi cũng chạy lại phương pháp deepMiRGene với source code các tác giả cung cấp, sau đó cũng đưa vào trong Bảng 2 Bảng 3 Bảng 4 để đánh giá. Các chỉ số như đã được được mô tả bên trên được tính như sau: TP: true positive, TN: true negative, FP: false positive, FN: false negative, SE = TP / (TP + FN), SP = TN / (TN + FP), PPV = TP / (TP + FP), F-score = 2TP / (2TP + FP + FN), g-mean = . Ngưỡng quyết định là 0.5.

Vì chúng ta đang xử lý một tập dữ liệu mất cân bằng (có rất nhiều mẫu “negative”), AUPR và AUROC rất quan trọng vì chúng không dễ bị mất cân bằng trong các lớp. AUPR đặc biệt quan trọng đối với chúng tôi vì mục tiêu chính của chúng tôi là phát hiện các mẫu “positive”. Độ chính xác đo lường xác suất phát hiện mẫu “positive” một cách chính xác và không bị ảnh hưởng bởi một số lượng lớn mẫu “negative” trong tập dữ liệu.

## Kết quả của 5-fold cross validation

Nhìn vào kết quả 5-fold cross validation, cách tiếp cận của chúng tôi đã vượt trội so với các phương pháp hiện đại trên bộ dữ liệu *human* và đạt được các kết quả cạnh tranh về tập dữ liệu *cross-spieces*. Cụ thể đối với bộ dữ liệu *human*, SE (0.878) tốt hơn 6% so với kết quả của phương pháp khác (0.818), trong khi F-score, g-mean, AUROC và AUPR thì cũng tốt hơn so với các phương pháp khác. Điều này cho thấy rằng cách tiếp cận của chúng tôi có thể xử lý dữ liệu mất cân bằng và tăng hy vọng cho kết quả tốt trong giai đoạn thử nghiệm (testing).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Methods | SE | SP | PPV | F-score | | g-mean | AUROC | AUPR |
| miRBoost (CV) | 0.803 | 0.988 | 0.887 | 0.843 | 0.891 | | - | - |
| CSHMM (CV) | 0.713 | 0.777 | 0.559 | 0.570 | 0.673 | | - | - |
| triplet-SVM (CV) | 0.669 | 0.986 | 0.851 | 0.749 | 0.812 | | 0.957 | 0.854 |
| microPred (CV) | 0.763 | **0.989** | **0.888** | 0.820 | 0.869 | | 0.974 | 0.890 |
| MIReNA (CV) | 0.818 | 0.943 | 0.624 | 0.708 | 0.878 | | - | - |
| LSTM-ATT-paper (CV) | 0.799 | 0.988 | 0.885 | 0.839 | 0.888 | | 0.984 | 0.915 |
| LSTM-ATT-rerun (CV) | 0.818 | 0.977 | 0.806 | 0.811 | 0.894 | | 0.975 | 0.876 |
| Proposed fixed-sized (CV) | **0.878** | 0.978 | 0.827 | 0.849 | **0.926** | | 0.984 | 0.915 |
| Proposed variable-sized (CV) | 0.835 | 0.985 | 0.868 | **0.851** | 0.907 | | **0.985** | **0.922** |
| Without contact map (CV) | 0.722 | 0.938 | 0.594 | 0.646 | 0.823 | | 0.909 | 0.726 |
| miRBoost (test) | 0.884 | 0.969 | 0.768 | 0.822 | 0.925 | | - | - |
| CSHMM (test) | 0.616 | 0.978 | 0.768 | 0.684 | 0.777 | | - | - |
| triplet-SVM (test) | 0.744 | **0.992** | 0.914 | 0.821 | 0.859 | | 0.947 | 0.830 |
| microPred (test) | 0.779 | 0.988 | 0.882 | 0.827 | 0.877 | | 0.980 | 0.892 |
| MIReNA (test) | 0.826 | 0.941 | 0.617 | 0.706 | 0.881 | | - | - |
| LSTM-ATT-paper (test) | 0.822 | **0.992** | **0.919** | 0.868 | 0.903 | | 0.981 | 0.918 |
| LSTM-ATT-rerun (test) | 0.767 | 0.989 | 0.892 | 0.825 | 0.871 | | 0.981 | 0.900 |
| Proposed fixed-sized (test) | **0.930** | 0.984 | 0.870 | **0.899** | **0.957** | | 0.983 | **0.946** |
| Proposed variable-sized (test) | 0.884 | 0.991 | 0.916 | **0.899** | 0.936 | | **0.986** | 0.934 |
| Without contact map (test) | 0.628 | 0.956 | 0.621 | 0.624 | 0.775 | | 0.892 | 0.669 |

Bảng 2. Kết quả trên tập dữ liệu human

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Methods | SE | SP | PPV | F-score | g-mean | AUROC | AUPR |
| miRBoost (CV) | 0.861 | 0.977 | 0.884 | 0.872 | 0.917 | - | - |
| CSHMM (CV) | 0.826 | 0.576 | 0.533 | 0.564 | 0.524 | - | - |
| triplet-SVM (CV) | 0.735 | 0.967 | 0.819 | 0.775 | 0.843 | 0.943 | 0.869 |
| microPred (CV) | 0.825 | 0.975 | 0.875 | 0.848 | 0.897 | 0.970 | 0.873 |
| MIReNA (CV) | 0.766 | 0.952 | 0.765 | 0.765 | 0.854 | - | - |
| LSTM-ATT-paper (CV) | 0.886 | **0.982** | **0.911** | **0.898** | 0.933 | **0.985** | 0.927 |
| LSTM-ATT-rerun (CV) | 0.883 | 0.970 | 0.861 | 0.871 | 0.926 | 0.981 | 0.929 |
| Proposed fixed-sized (CV) | **0.903** | 0.978 | 0.894 | **0.898** | **0.940** | **0.985** | **0.936** |
| Proposed variable-sized (CV) | 0.881 | 0.981 | 0.906 | 0.893 | 0.930 | 0.983 | **0.936** |
| Without contact map (CV) | 0.785 | 0.928 | 0.704 | 0.737 | 0.853 | 0.926 | 0.828 |
| miRBoost (test) | 0.856 | 0.844 | 0.526 | 0.651 | 0.850 | - | - |
| CSHMM (test) | 0.749 | 0.960 | 0.791 | 0.769 | 0.848 | - | - |
| triplet-SVM (test) | 0.760 | 0.977 | 0.870 | 0.812 | 0.862 | 0.952 | 0.908 |
| microPred (test) | 0.814 | 0.985 | 0.919 | 0.863 | 0.896 | 0.963 | 0.906 |
| MIReNA (test) | 0.796 | 0.950 | 0.764 | 0.780 | 0.870 | - | - |
| LSTM-ATT-paper (test) | 0.900 | 0.983 | 0.913 | 0.906 | 0.940 | 0.984 | **0.955** |
| LSTM-ATT-rerun (test) | 0.862 | 0.984 | 0.917 | 0.889 | 0.921 | 0.984 | 0.954 |
| Proposed fixed-sized (test) | **0.904** | 0.982 | 0.910 | **0.907** | **0.942** | 0.983 | 0.951 |
| Proposed variable-sized (test) | 0.880 | **0.988** | **0.936** | **0.907** | 0.933 | **0.985** | 0.950 |
| Without contact map (test) | 0.695 | 0.955 | 0.758 | 0.725 | 0.815 | 0.914 | 0.798 |

Bảng 3. Kết quả trên tập dữ liệu cross-species

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Methods | SE | SP | PPV | F-score | g-mean | AUROC | AUPR |
| miRBoost | **0.921** | 0.936 | 0.609 | 0.733 | 0.928 | - | - |
| CSHMM | 0.536 | 0.069 | 0.046 | 0.085 | 0.192 | - | - |
| triplet-SVM | 0.721 | **0.981** | 0.759 | 0.740 | 0.841 | 0.934 | 0.766 |
| microPred | 0.728 | 0.970 | 0.672 | 0.699 | 0.840 | 0.940 | 0.756 |
| MIReNA | 0.450 | 0.941 | 0.392 | 0.419 | 0.650 | - | - |
| LSTM-ATT-paper | 0.917 | 0.964 | 0.682 | 0.782 | 0.941 | **0.981** | 0.808 |
| LSTM-ATT-rerun | 0.899 | 0.968 | 0.700 | 0.787 | 0.933 | 0.980 | 0.798 |
| Proposed fixed-sized | 0.917 | 0.967 | 0.696 | 0.792 | **0.942** | 0.797 | **0.864** |
| Proposed variable-sized | 0.859 | **0.981** | **0.779** | **0.817** | 0.918 | 0.979 | 0.818 |
| Without contact map | 0.855 | 0.938 | 0.535 | 0.658 | 0.896 | 0.958 | 0.798 |

Bảng 4. Kết quả trên tập dữ liệu new

## Kết quả trên dữ liệu test

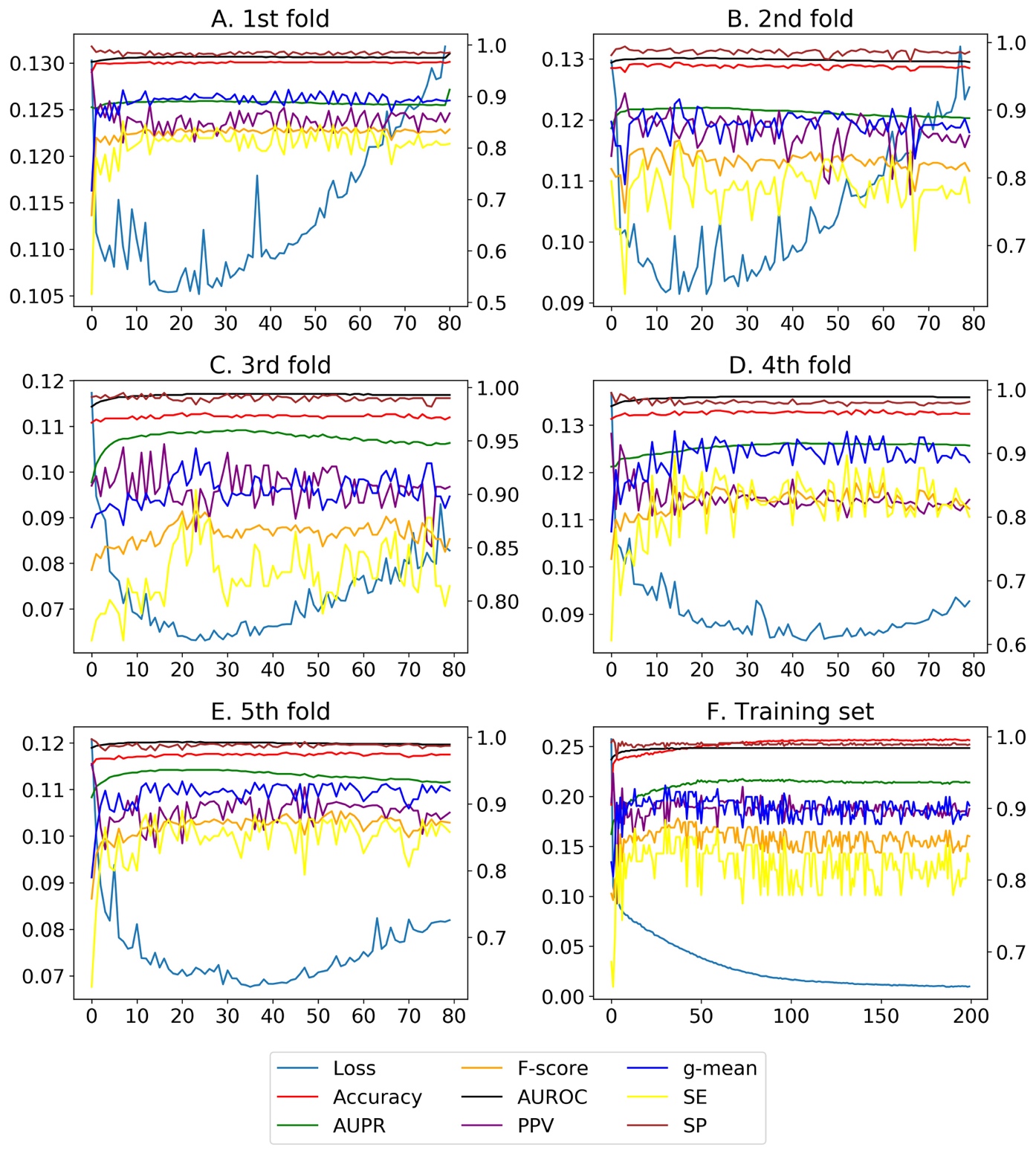
Phương pháp được đề xuất tốt hơn các phương pháp trước khá ổn định. Trong bộ dữ liệu *human*, kết quả kiểm tra của mỗi số liệu (ngoại trừ SP và PPV) đều tốt hơn tất cả các phương pháp khác. Đối với tập dữ liệu *cross-spieces*, phương pháp của chúng tôi cho thấy các kết quả có thể so sánh với các phương pháp khác về SP, PPV và F-score và AUROC. Trên tập dữ liệu *new*, nó đạt được điểm số cao nhất về SP, PPV, F-score và AUPR. Ngoài ra, trong tất cả các tập dữ liệu, chúng tôi đạt được điểm số F-score (0.899 cho *human*, 0.907 đối với *cross-spieces* và 0.817 đối với *new*). Ngay cả trong các số liệu mà phương pháp của chúng tôi không có điểm số tốt nhất, nó là rất gần với tốt nhất trong hầu như tất cả các trường hợp. Trong các bộ dữ liệu *human* và *new*, có thể thấy rằng phương pháp của chúng tôi một sự cải thiện đáng kể so với deepMiRGene.

Chúng ta cũng có thể thấy cách tiếp cận học tập sâu mà không có các đặc trưng thủ công (handcrafted features), cụ thể là deepMiRGene và mô hình của chúng tôi, tốt hơn so với các phương pháp học máy trước đó. Các phương pháp mới có kết quả tốt hơn trong tập dữ liệu *new* trong mỗi số liệu ngoại trừ tính SP, trong đó triplet-SVM cho cùng một điểm như deepMiRGene. Đối với cả hai phần cross validation và test của *cross-spieces*, các phương pháp học sâu đưa ra kết quả tốt hơn tất cả các phương pháp trước đây. Đối với tập dữ liệu *new*, miRBoost cho kết quả tốt nhất về mặt SE, và triplet-SVM có cùng SP như mô hình được đề xuất của chúng tôi, nhưng trong mọi phương diện khác, phương pháp học sâu có kết quả tốt hơn. Xem xét tất cả các tập dữ liệu và kết quả, chúng ta có thể kết luận rằng phương pháp học sâu tốt hơn các giải pháp học máy trước đây.

Khi so sánh deepMiRGene với phương pháp của chúng tôi, cách tiếp cận của chúng tôi cho kết quả tốt hơn trong tất cả các tập dữ liệu thử nghiệm. Đối với tập dữ liệu *human*, cách tiếp cận của chúng tôi cung cấp các cải tiến về mọi chỉ số ngoại trừ SP và PPV. Đối với tập dữ liệu *cross-spieces*, phương pháp của chúng tôi cho kết quả tốt hơn về mọi số liệu khác, ngoại trừ AUPR. Cuối cùng, đối với tập dữ liệu *new*, mô hình của chúng tôi hoạt động tốt hơn trong mọi chỉ số ngoại trừ AUROC. Kết quả khi training trên tập dữ liệu *cross-spieces* và test trên tập dữ liệu *new* (Bảng 4) chứng minh rằng cách tiếp cận của chúng tôi có thể xác định tốt các pre-miRNA trong một loài mới mặc dù nó được huấn luyện trên một tập hợp các loài hỗn hợp.

Cũng quan trọng để so sánh hai cách tiếp cận của chúng tôi với nhau. Nếu không có ma trận ghép nối (pairing matrix), kết quả sẽ tồi tệ hơn nhiều trong mỗi tập dữ liệu và mỗi số liệu. Do đó, ma trận ghép nối là một phần quan trọng trong phương pháp của chúng tôi. Các đầu vào có kích thước cố định và có kích thước biến thiên cho kết quả tương tự cho hầu hết dữ liệu. Với đầu vào có kích thước cố định, chúng tôi cần nhiều bộ nhớ để huấn luyện hơn so với đầu vào có kích thước thay đổi. Do đó, chúng tôi khuyến khích sử dụng đầu vào có kích thước biến thiên cho nhiệm vụ này.

Xem xét tất cả các kết quả và so sánh với các phương pháp khác, kết quả cho thấy phương pháp đề xuất của chúng tôi mang lại hiệu suất tốt nhất trong việc xác định các pre-miRNA.



Hình 11. Biểu đồ các chỉ số đánh giá trên bộ dữ liệu 5-fold và training

# Kết luận và công việc tương lai

Kết quả khi sử dụng *pairing* matrix tốt hơn nhiều so với khi không sử dụng nó. Điều đó có nghĩa rằng *pairing matrix* và cấu trúc bậc hai đóng vai trò quan trọng trong nhận biết miRNA. Bằng cách nhúng cấu trúc bậc hai vào đầu vào của mạng ConvNet, không gian đặc trưng sẽ lớn hơn. Vì thế, chúng ta cần một bộ dữ liệu lớn hơn để tổng quát hoá được dữ liệu. Bên cạnh đó, chúng ta phải đối mặt với sự mất cân bằng giữa các nhãn (positive và negative) trong dữ liệu. Một hướng hứa hẹn cho công việc tương lai để cải thiện kết quả nhận biết là sưu tập thêm nhiều dữ liệu, sau đó sử dụng mạng nơron lớn hơn, chẳng hạn AlexNet [43] hoặc ResNet [38].

Trong đồ án này, em đã đề xuất cách thể hiện thông tin cấu trúc bậc hai của RNA, hay còn gọi là *pairing matrix,* làm đầu vào cho một mạng ConvNet để trích xuất các đặc trưng một cách tự động. Phương pháp này đã giành được kết quả tốt nhất trên ba bộ dữ liệu chuẩn *human*, *cross-species* and new, đặc biệt cho độ đo F-score. Trong tương lai, em sẽ mở rộng phương pháp để giải quyết các bài toán khác liên quan đến miRNA, chẳng hạn: dự đoán chuỗi mRNA đích của miRNA, dự đoán chức năng của miRNA [49].

# Tài liệu Tham Khảo

|  |  |
| --- | --- |
| [1] | B. Isaac, “Identification of hundreds of conserved and nonconserved human microRNAs,” *Nature Genetics,* tập 37, p. 766, 2005. |
| [2] | M. Jones-Rhoades, “MicroRNAs and their regulatory roles in plants,” *Annu. Rev. Plant Biol.,* tập 57, pp. 19-53, 2006. |
| [3] | S. Pfeffer, “Identification of virus-encoded microRNAs,” *Science,* tập 304, pp. 734-736, 2004. |
| [4] | D. P. Bartel, “MicroRNAs: genomics, biogenesis, mechanism, and function,” *Cell,* tập 116, pp. 281-297, 2004. |
| [5] | M. D. Jansson, “MicroRNA and cancer,” *Molecular oncology,* tập 6, pp. 590-610, 2012. |
| [6] | D. Wang, “Human microRNA oncogenes and tumor suppressors show significantly different biological patterns: from functions to targets,” *PloS one,* tập 5, 2010. |
| [7] | Tüfekci, “The role of microRNAs in human diseases,” *miRNomics: MicroRNA Biology and Computational Analysis,* pp. 33-50, 2014. |
| [8] | A. Mathelier, “MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data,” *Bioinformatics,* tập 26, pp. 2226-2234, 2010. |
| [9] | A. Gudyś, “HuntMi: an efficient and taxon-specific approach in pre-miRNA identification,” *BMC bioinformatics,* tập 14, p. 83, 2013. |
| [10] | S. Tempel, “miRBoost: boosting support vector machines for microRNA precursor classification,” *RNA,* 2015. |
| [11] | S. Agarwal, “Prediction of novel precursor miRNAs using a context-sensitive hidden Markov model (CSHMM),” *BMC bioinformatics,* tập 11, p. S29, 2010. |
| [12] | R. Batuwita, “microPred: effective classification of pre-miRNAs for human miRNA gene prediction,” *Bioinformatics,* tập 25, pp. 989-995, 2009. |
| [13] | P. Jiang, “MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features,” *Nucleic acids research,* tập 35, số W339-W344, p. Oxford University Press, 2007. |
| [14] | C. Xue, “Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine,” *BMC bioinformatics,* tập 6, số BioMed Central, p. 310, 2005. |
| [15] | M. E. Rahman, “MiRANN: A reliable approach for improved classification of precursor microRNA using Artificial Neural Network model,” *Genomics,* tập 99, số Elsevier, pp. 189-194, 2012. |
| [16] | J. Thomas, “DP-miRNA: An improved prediction of precursor microRNA using deep learning model,” *Big Data and Smart Computing (BigComp), 2017 IEEE International Conference,* số IEEE, pp. 96-99, 2017. |
| [17] | S. Park, “Deep recurrent neural network-based identification of precursor micrornas,” *Advances in Neural Information Processing Systems,* pp. 2891-2900, 2017. |
| [18] | I. L. Hofacker, “Vienna RNA secondary structure server,” *Nucleic acids research,* tập 31, số Oxford University Press, pp. 3429-3431, 2003. |
| [19] | A. Mathuriya, “GTfold: a scalable multicore code for RNA secondary structure prediction,” *Proceedings of the 2009 ACM symposium on Applied Computing,* số ACM, pp. 981-988, 2009. |
| [20] | E. Bindewald, “Cylofold: secondary structure prediction including pseudoknots,” *Nucleic acids research,* tập 38, số Oxford University Press, pp. W368-W372, 2010. |
| [21] | L. Jiang, “BP neural network could help improve pre-miRNA identification in various species,” *BioMed research international,* số Hindawi, 2016. |
| [22] | K. L. S. Ng, “De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures,” *Bioinformatics,* tập 23, số Oxford University Press, pp. 1321-1330, 2007. |
| [23] | V. D. T. Tran, “miRBoost: boosting support vector machines for microRNA precursor classification,” *RNA,* tập 21, p. 775. |
| [24] | Demirci, “ On the performance of pre-microRNA detection algorithms,” *Nature Communications,* tập 8, p. 330. |
| [25] | J. Kuka{\v{c}}ka, “Regularization for Deep Learning: A Taxonomy,” *arXiv,* 2017. |
| [26] | L. Cun, “Deep learning,” *Nature,* tập 521, p. 436. |
| [27] | J. Mendenhall, “Improving quantitative structure--activity relationship models using Artificial Neural Networks trained with dropout,” *Journal of computer-aided molecular design,* tập 30, số Springer, pp. 177-189, 2016. |
| [28] | V. Golkov, “3D deep learning for biological function prediction from physical fields,” *arXiv,* 2017. |
| [29] | M. Spencer, “A deep learning network approach to ab initio protein secondary structure prediction,” *IEEE/ACM transactions on computational biology and bioinformatics (TCBB),* tập 12, pp. 103-112, 2015. |
| [30] | M. AlQuraishi, “End-to-end differentiable learning of protein structure,” *bioRxiv,* số Cold Spring Harbor Laboratory, 2018. |
| [31] | V. Golkov, “Protein contact prediction from amino acid co-evolution using convolutional networks for graph-valued images,” *Advances in Neural Information Processing Systems,* pp. 4222-4230, 2016. |
| [32] | S. Hochreiter, “Long short-term memory,” *Neural computation,* tập 9, pp. 1735-1780, 1997. |
| [33] | A. Churkin, “RNA dot plots: an image representation for RNA secondary structure analysis and manipulations,” *Wiley Interdisciplinary Reviews: RNA,* tập 4, số Wiley Online Library, pp. 205-216, 2013. |
| [34] | W. M. Fitch, “Locating gaps in amino acid sequences to optimize the homology between two proteins,” *Biochemical genetics,* tập 3, pp. 99-108, 1969. |
| [35] | J. V. Maizel, “Enhanced graphic matrix analysis of nucleic acid and protein sequences,” tập 78, pp. 7665-7669, 1981. |
| [36] | R. Staden, “An interactive graphics program for comparing and aligning nucleic acid and amino acid sequences,” *Nucleic Acids Research,* tập 10, số Oxford University Press, pp. 2951-2961, 1982. |
| [37] | X. Wang, “miRDB: a microRNA target prediction and functional annotation database with a wiki interface.,” *RNA,* tập 14, số Cold Spring Harbor Lab, pp. 1012-1017, 2008. |
| [38] | K. He, "Deep residual learning for image recognition," in *the IEEE conference on computer vision and pattern recognition*, 2016. |
| [39] | A. Singh, “Sentiment Transfer using Seq2Seq Adversarial Autoencoders,” *Arxiv,* 2018. |
| [40] | G. H. Weinberger, “Densely Connected Convolutional Networks,” trong *Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017. |
| [41] | J. Duchi, “Adaptive subgradient methods for online learning and stochastic optimization,” *Journal of Machine Learning Research,* tập 12, pp. 2121-2159, 2011. |
| [42] | D. Kingma, “A method for stochastic optimization,” *arXiv,* 2014. |
| [43] | A. Krizhevsky, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012. |
| [44] | N. Srivastava, “Dropout: a simple way to prevent neural networks from overfitting,” *The Journal of Machine Learning Research,* tập 15, pp. 1929-1958, 2014. |
| [45] | A. Krogh, “A simple weight decay can improve generalization,” *Advances in neural information processing systems,* pp. 950-957, 1992. |
| [46] | Griffiths-Jones, “miRBase: microRNA sequences, targets and gene nomenclature,” *Nucleic acids research,* tập 34, số Oxford University Press, pp. D140-D144, 2006. |
| [47] | D. Bu, “NONCODE v3. 0: integrative annotation of long noncoding RNAs,” *Nucleic acids research,* tập 40, số Oxford University Press, pp. D210-D215, 2011. |
| [48] | L. Lestrade, “snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs,” *Nucleic acids research,* tập 34, số Oxford University Press, pp. D158-D162, 2006. |
| [49] | S. Singh, "Machine Learning Techniques in Exploring MicroRNA Gene Discovery, Targets, and Functions," *Bioinformatics in MicroRNA Research,* no. Springer, pp. 211-224, 2017. |