ỨNG DỤNG MULTITASK-LEARNING TRONG PHÂN TÍCH THÀNH PHẦN HÓA HỌC THỰC PHẨM DỰA TRÊN PHỔ NIR

SVTH: Nguyễn Huy Tưởng, Lê Hoàng Ngọc Hân

Lớp 19TCLC_DT4, Khoa Công nghệ Thông tin, Trường Đại học Bách Khoa - Đại học Đà Nẵng;
Email: huytuong010101@gmail.com, hanlehngoc080601@gmail.com

GVHD: TS. Nguyễn Văn Hiệu, KS. Lưu Văn Huy

Khoa Công nghệ Thông tin, Trường Đại học Bách Khoa - Đại học Đà Nẵng;
Email: nvhieuqt@gmail.com

Tóm tắt - Việc dự đoán thành phần hóa học của thực phẩm là cực kỳ quan trọng, điều này giúp chúng ta đánh giá được mức độ an toàn và mức độ dinh dưỡng của của thực phẩm. Trong bài nghiên cứu này, chúng tôi đề xuất mô hình kết hợp giữa Multitask Learning và Neural Network để có thể phân tích thành phần hóa học của nhiều loại thực phẩm khác nhau trong cùng một mô hình dựa trên phổ NIR. Việc sử dụng phổ cận hồng ngoại (NIR) giúp việc nhận diện và phân tích các thành phần trong thực phẩm nhanh chóng và đơn giản, đồng thời không làm ảnh hưởng đến chất lượng của thực phẩm. Trong bài nghiên cứu này chúng tôi đề xuất một mô hình end-to-end có khả năng phân tích thành phần hóa học của nhiều chất của nhiều loài thực phẩm cùng lúc. Kết quả tương đối khả quan cho thấy khả năng đây sẽ là giải pháp công nghệ thông tin thích hợp để hỗ trợ kiểm soát chất lượng thực phẩm trên thị trường.

Từ khóa - Thành phần hóa học, phổ cận hồng ngoại, multitask learning.

Abstract - Predicting the chemical composition of food is extremely important, which helps us assess the safety and nutritional level of food. In this article, we propose a combine of Multi-task Learning and Neural Network which allows us to analyze chemical composition of different kinds of food in only one model based on NIR. Applying NIR helps food classification and chemical substances analysis become faster and much simple, and at the same time does not affect the quality of the food. The end-to-end model we propose in this article is capable of analyzing multiple substances of multiple kinds of food at the same time. The obtained results show the possibility of a proper technological solution for controlling food quality in the market.

Key words - chemical composition, chemical substances analysis, near infra-red, multi-task learning.

1. Đặt vấn đề

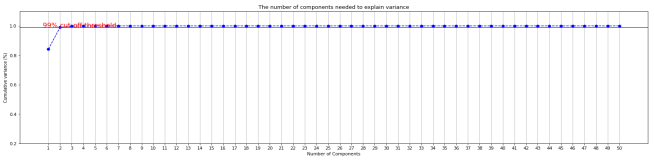
Hiện nay, an toàn thực phẩm đang là vấn đề nhức nhối trong xã hội. Với tình trạng thực phẩm bẩn, chứa các chất độc hại vẫn đang trôi nổi trên thị trường hiện nay đặt ra vấn đề cần có một phương pháp để giúp đánh giá chất lượng của thực phẩm dựa trên thành phần hóa học của nó. Phương pháp phổ biến hiện này là thực hiện phân tích hóa học trong phòng thí nghiệm Tuy nhiên phương pháp này đòi hỏi tốn nhiều thời gian, công sức, khó áp dụng, đồng thời điểm quan trọng nhất là chúng ta cần tác động vật lý đến thực phẩm đó để lấy mẫu phân tích.

Hiện nay, với sự phát triển của khoa học và công nghệ, quang phổ cận hồng ngoại đang là phương pháp hiệu quả nhất để phân tích thành phần hóa học của một chất. Các nghiên cứu trên các mô hình học máy như Hồi quy tuyến tính, Hồi quy bình phương tối thiểu từng phần, giảm chiều dữ liệu (PCA, LDA,...) cho thấy rằng dựa vào phổ NIR có thể phân tích được thành phần hóa học của các loại thực phẩm. Tuy nhiên các nghiên cứu này chỉ được thực hiện và tối ưu hóa trên cùng một loại thực phẩm. Điều này đặt ra vấn đề khi áp dụng vào thực tế, cần phân tích thành phần hóa học của nhiều loại thực phẩm khác nhau, nếu ứng dụng các phương pháp này thì cần phải đi phân tích và tối ưu hóa trên từng loại thực phẩm. Dựa trên các kết quả nghiên cứu đó, chúng tôi đề xuất sử dụng Neural Network kết hợp với kỹ thuật MultiTask Learning để xây dựng một mô hình có khả năng dự đoán thành phần hóa học của nhiều chất khác nhau. Từ đó ứng dụng vào các hệ thống thực tế giúp kiểm soát chất lượng thực phẩm.

2. Phân tích dữ liệu

2.1. Dữ liệu sử dụng

Trong bài nghiên cứu này, chúng tôi sử dụng tập dữ liệu [1] bao gồm 186 phổ NIR được tác giả thu thập từ 4 giống xoài khác nhau với bước sóng từ 1000 - 2500nm. Dữ liệu quang phổ này có khả năng được sử dụng và phân tích để dự đoán các thuộc tính chất lượng của xoài dưới dạng vitamin C, hàm lượng chất rắn hòa tan (SSC) và tổng lượng axit (TA).

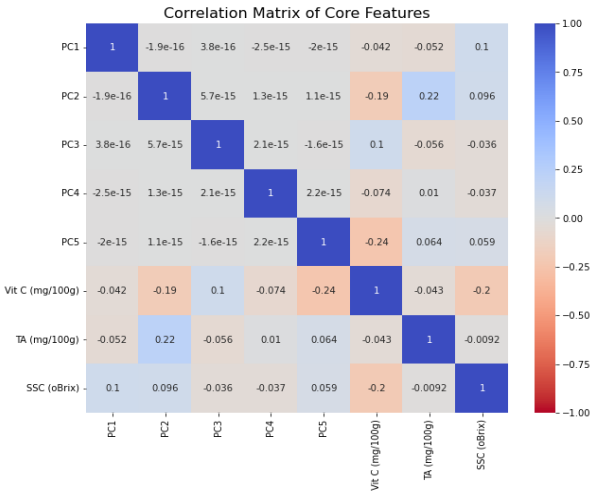


Hình x. Khảo sát số chiều PCA có thể giảm về đối với giá trị phổ ứng với dải bước sóng từ 1000 - 2500nm

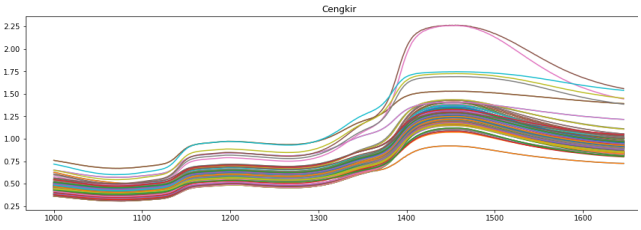
Có thể thấy chỉ cần giảm về số chiều ~ 3 thì độ thông tin của dữ liệu đã đạt hơn 99%. Để đảm bảo độ tin cậy, nhóm chọn số chiều là 5 để khảo sát độ tương quan giữa dữ liệu phổ với nồng độ các chất có trong quả.

2.2. Đặc trưng phổ của từng loại xoài

2.2.1. Xoài Cengkir

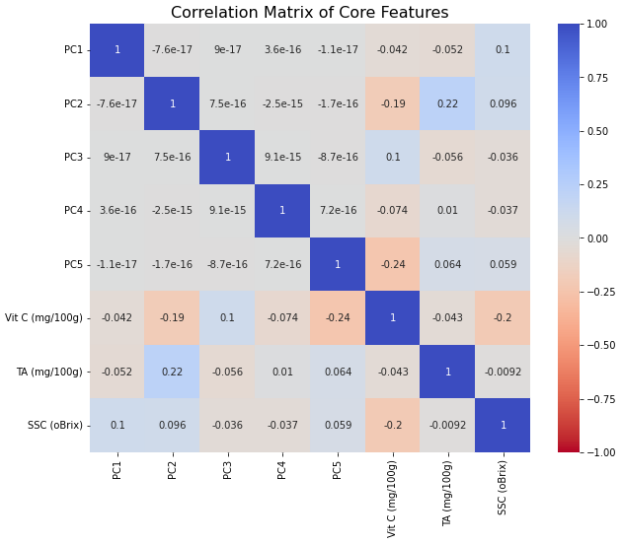


Hình 1. Ma trận tương quan giữa phổ tại các bước sóng (sau khi giảm chiều PCA) với nồng độ các chất có trong xoài Cengkir

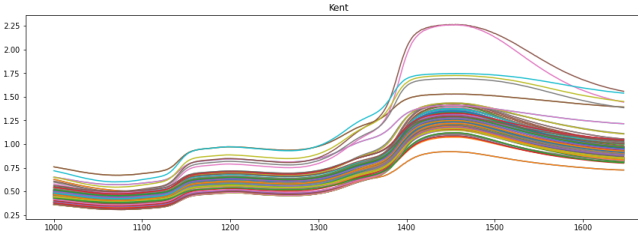


Hình 2. Dải phổ theo bước sóng của xoài Cengkir

2.2.2. *Xoài Kent*

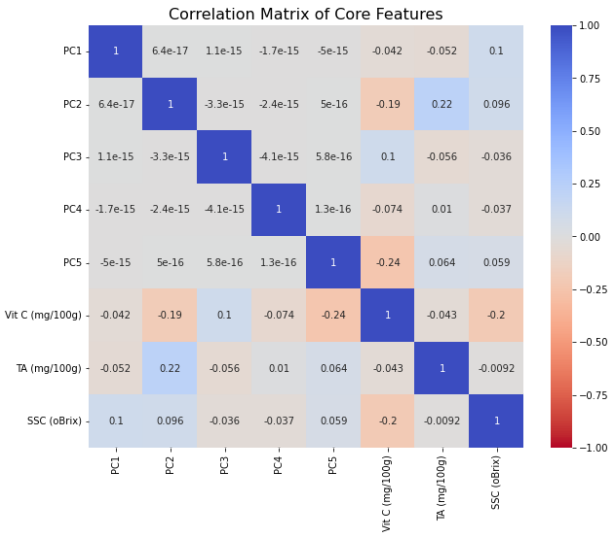


Hình 3. Ma trận tương quan giữa phổ tại các bước sóng (sau khi giảm chiều PCA) với nồng độ các chất có trong xoài Kent

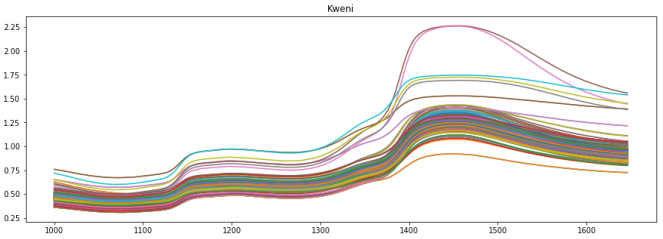


Hình 4. Dải phổ theo bước sóng của xoài Kent

2.2.3. *Xoài Kweni*

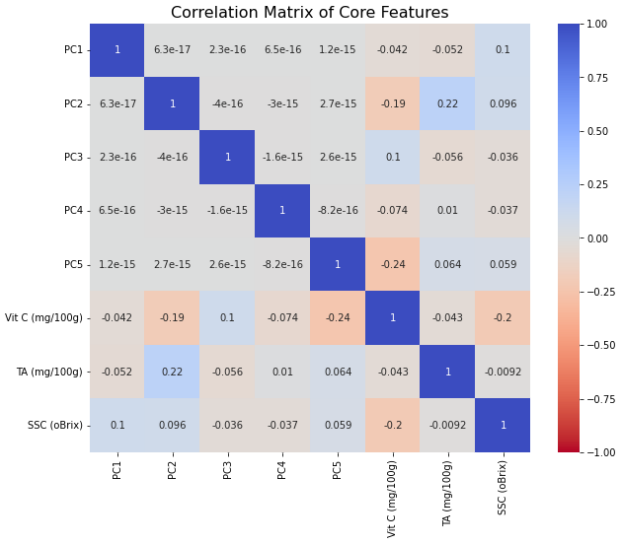


Hình 5. Ma trận tương quan giữa phổ tại các bước sóng (sau khi giảm chiều PCA) với nồng độ các chất có trong xoài Kweni

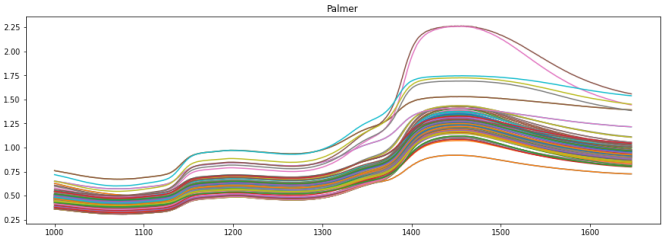


Hình x. Dải phổ theo bước sóng của xoài Kweni

2.2.4. *Xoài Palmer*



Hình 6. Ma trận tương quan giữa phổ tại các bước sóng (sau khi giảm chiều PCA) với nồng độ các chất có trong xoài Palmer



Hình 7. Dải phổ theo bước sóng của xoài Palmer

3. Kết quả nghiên cứu và khảo sát

3.1. Các phương pháp tiền xử lý dữ liệu

3.1.1. Xử lý dữ liệu trống

Dữ liệu trong quá trình thu thập có thể bị khuyết hoặc mất đi. Nguyên nhân có thể phát sinh từ nhiều yếu tố như: thiết bị đo gặp lỗi trong một thời điểm, quá trình truyền dữ liệu từ thiết bị đo tới nơi lưu trữ gặp vấn đề,...Dưới đây là một số phương pháp xử lý dữ liệu trống mà nhóm đã tìm hiểu, nghiên cứu và thử nghiệm.

a. Loại bỏ dữ liệu trống

Dữ liệu phổ NIR được đo tại từng bước sóng cụ thể. Thông thường, dữ liệu hay bị trống hàng loạt ở một vài bước sóng nhất định, trong trường hợp này ta có thể giải quyết bằng cách bỏ đi cột giá trị tại bước sóng đó. Tuy nhiên, phương pháp xử lý dữ liệu trống bằng việc loại bỏ thường không được khuyến khích và chỉ thực hiện bất khả thi vì nó sẽ làm dữ liệu mất đi thông tin và độ tin cậy.

b. Thay thế dữ liệu trống bằng các giá trị khác

Phương pháp thay thế này có thể khắc phục được nhược điểm của phương pháp loại bỏ giá trị trống. Tuy nhiên, chỉ thực hiện được khi giá trị bị trống ít, thưa thớt hoặc dữ liệu có quy luật có thể suy ngược lại.

Một số phương pháp thay thế đơn giản có thể áp dụng với dữ liệu trống ít: Thay thế dữ liệu trống bằng giá trị trung bình, trung vị, mode (giá trị xuất hiện thường xuyên nhất, giá trị 0, …). Các phương pháp này được hỗ trợ bởi một số thư viện của Python (sklearn, pandas, numpy), Matlab, R,...

Tuy nhiên, các phương pháp đơn giản lại thường không mang lại hiệu quả cao và thậm chí có thể gây nhiễu cho mô hình dự đoán. Chính vì vậy mà nhóm đã tìm hiểu và nghiên cứu thêm một số phương pháp nâng cao và phức tạp hơn như: các phương pháp hồi quy đơn và đa biến, hồi quy tuyến tính và phi tuyến tính, …

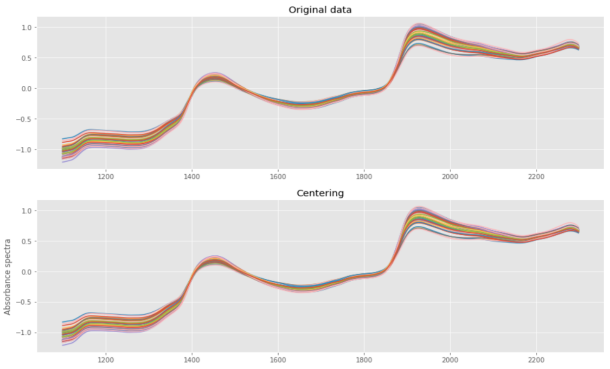
Hiện nay có một số thư viện hỗ trợ việc xử lý giá trị trống bằng cách phương pháp hồi quy như sklearn, scipy của Python,...

3.1.2. Chuẩn hóa dữ liệu

Chuẩn hóa dữ liệu giúp đưa dữ liệu về cùng một phân phối, điều chỉnh phạm vi của dữ liệu góp phần giảm chi phí tính toán và tính công bằng giữa các đặc trưng với nhau. Bên cạnh đó, chuẩn hóa cũng có nhiệm vụ xử lý các giá trị bất thường hoặc ngoại lệ có thể gây nhiễu, giúp cho mô hình có thể học được tốt hơn.

a. Phương pháp định tâm (Centering)

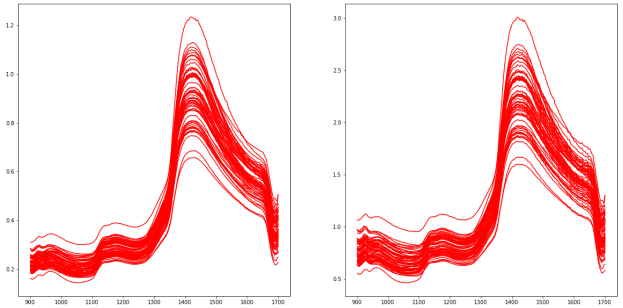
Phương pháp định tâm là cách biến đổi dữ liệu trong đó mỗi quan sát được trừ cho 1 giá trị nhất định(ví dụ giá trị trung bình của biến đó). Việc này tương tự như chuẩn hóa 1 biến standardization ngoại trừ không có chia cho giá trị độ lệch chuẩn. Mục tiêu để tăng cường khả năng so sánh giữa các biến. Kết quả minh họa phương pháp có trong Hình 8.



Hình 8. Hình dạng phổ ban đầu (trên) so với phổ lúc thực hiện phương pháp định tâm (dưới)

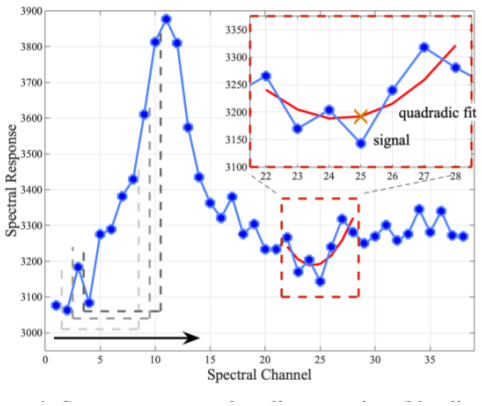
b. Phương pháp chuẩn hóa (Standardization)

Phương pháp chuẩn hóa là việc đưa dữ liệu về một phân bố trong đó giá trị trung bình của các quan sát bằng 0 và độ lệch chuẩn = 1. Kỹ thuật này còn được gọi là "whitening". Phương pháp này không làm thay đổi hình dạng phổ mà chỉ thay đổi dải phân phối giá trị của dữ liệu. Kết quả minh họa phương pháp có trong Hình 9.



Hình 9. Hình dạng phổ ban đầu (trái) so với phổ lúc thực hiện phương pháp chuẩn hóa (phải)

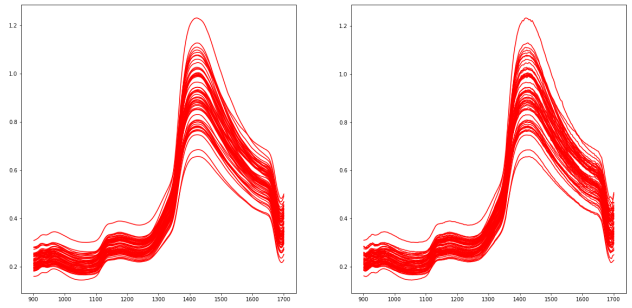
c. Phương pháp biến đổi (Transformation)



Hình 10. Minh họa phương pháp biến đổi dữ liệu

Phương pháp biến đổi dữ liệu là quá trình sửa đổi, tính toán, phân tách và kết hợp dữ liệu thô thành các mô hình dữ liệu sẵn sàng phân tích. Đối với bài toán phổ NIR, nhóm đã ứng dụng phương pháp này để có để làm mịn phổ, có nghĩa là lọc bớt những giá trị nhiễu đi để hình dạng phổ mượt hơn. Kết quả minh

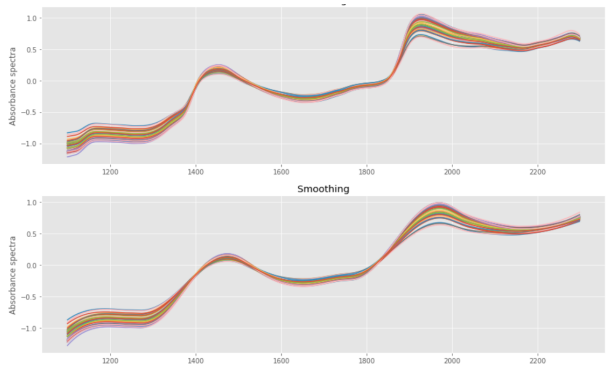
họa phương pháp có trong Hình 11.



Hình 11. Hình dạng phổ ban đầu (trái) so với phổ lúc thực hiện phương pháp biến đổi dữ liệu (phải)

d. Phương pháp lọc mịn (Smoothing filter)

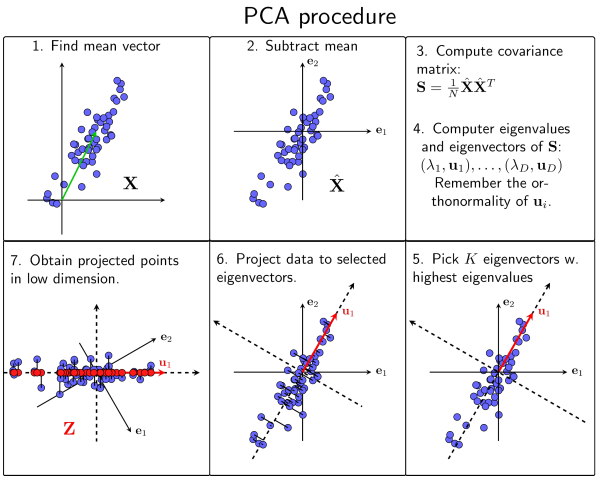
Phương pháp lọc mịn là phương thức xử lí dữ liệu, được thực hiện bằng cách sử dụng thuật toán để loại bỏ nhiễu khỏi bộ dữ liệu. Điều này cho phép các mẫu và xu hướng quan trọng trở nên nổi bật. Kết quả minh họa phương pháp có trên Hình 12



Hình 12. Hình dạng phổ ban đầu (trên) so với phổ lúc thực hiện phương pháp lọc mịn (dưới)

3.2. Thử nghiệm phân loại thực phẩm sử dụng mô hình học máy

3.2.1. *Phân tích thành phần chính (PCA)*



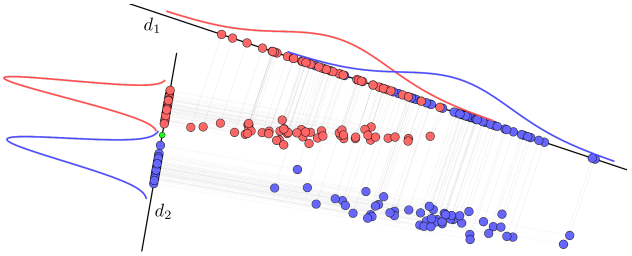
Hình 13. Minh họa thuật toán giảm chiều PCA

PCA là phương pháp đi tìm một hệ cơ sở mới sao cho thông tin của dữ liệu chủ yếu tập trung ở một vài toạ độ, phần còn lại chỉ mang một lượng nhỏ thông tin. Và để cho đơn giản trong tính toán, PCA sẽ tìm một hệ trực chuẩn để làm cơ sở mới.

Tuy nhiên, PCA là phương pháp thuộc loại học không giám sát, tức là nó chỉ sử dụng các vector mô tả dữ liệu mà không dùng tới nhãn, nếu có, của dữ liệu. Trong bài toán phân loại mà nhóm đang hướng đến, dạng điển hình nhất của học có giám sát, việc sử dụng nhãn sẽ mang lại kết quả phân loại tốt hơn.

3.2.2. *Phân tích phân biệt tuyến tính (LDA)*

Như đã đề cập, PCA là phương pháp giảm chiều dữ liệu sao cho lượng thông tin về dữ liệu, thể hiện ở tổng phương sai, được giữ lại là nhiều nhất. Tuy nhiên, trong nhiều trường hợp, ta không cần giữ lại lượng thông tin lớn nhất mà chỉ cần giữ lại thông tin cần thiết cho riêng bài toán, cụ thể ở đây là bài toán phân loại.



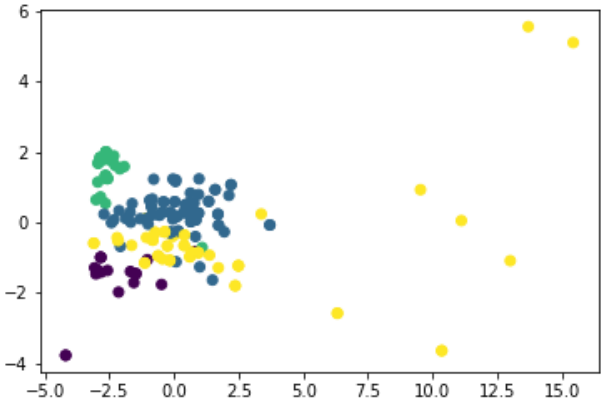
Hình 13. Minh họa việc classification đơn giản nhất có thể được hiểu là việc tìm ra một ngưỡng giúp phân tách hai class một cách đơn giản và đạt kết quả tốt nhất.

Dựa vào hình minh họa trên ta thấy, không phải việc giữ lại thông tin nhiều nhất sẽ luôn mang lại kết quả tốt nhất.

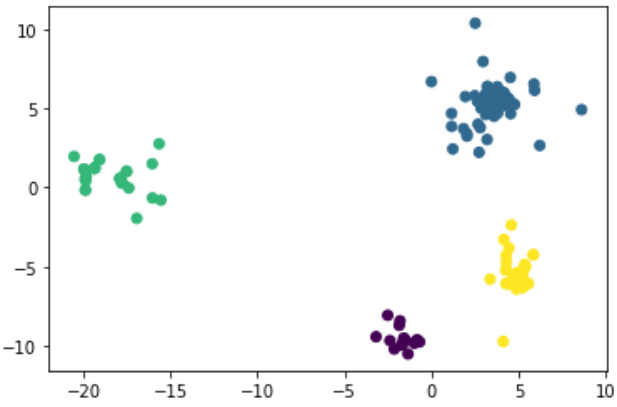
Phân tích phân biệt tuyến tính (LDA) ra đời nhằm giải quyết vấn đề này. LDA là một phương pháp giảm chiều dữ liệu cho bài toán phân lớp. LDA có thể được coi là một phương pháp giảm chiều dữ liệu, và cũng có thể được coi là một phương pháp phân lớp, và cũng có thể được áp dụng đồng thời cho cả hai, tức giảm chiều dữ liệu sao cho việc phân lớp hiệu quả nhất.

3.2.3. *Kết quả thực nghiệm*

Quá trình thực nghiệm được tiến hành trên tập dữ liệu nêu trên bao gồm 4 loại xoài: Cengkir (18 mẫu), Kent (85 mẫu), Kweni (29 mẫu) và Palmer (54 mẫu). Nhóm chia tập huấn luyện và kiểm thử với tỉ lệ 75:25.



Hình 14. Các điểm dữ liệu sau khi giảm chiều bằng PCA.



Hình 15. Các điểm dữ liệu sau khi giảm chiều bằng LDA.

Có thể thấy dữ liệu sau khi giảm chiều bằng PCA tuy có thể dữ liệu nhiều thông tin hơn nhưng hiệu quả phân lớp lại không tốt. Sử dụng LDA đem lại kết quả phân lớp vượt trội rõ rệt.

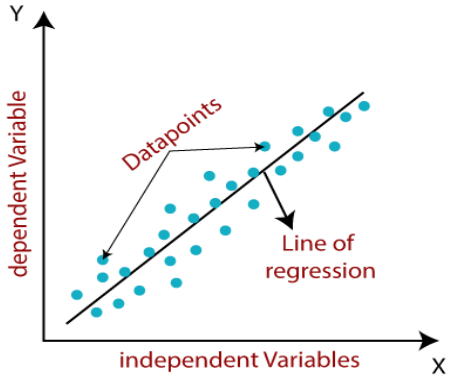
***Bảng 1**. Accuracy phân lớp trên tập dữ liệu khi sử dụng LDA và PCA kết hợp LDA.*

STT	Mô hình	Accuracy
1	LDA	0.96
2	PCA + LDA	1.0

3.3. Thử nghiệm dự đoán nồng độ các chất trong thực phẩm sử dụng mô hình học máy

3.3.1. Hồi quy tuyến tính

Hồi quy tuyến tính là một thuật toán học máy cơ bản, trong đó các biến đích (y) sẽ phụ thuộc vào sự biến thiên của các biến độc lập (X). Mô hình hoạt động với hàm số tuyến tính cơ bản: y = wX + b. Trong đó, quá trình huấn luyện mô hình là đi tìm bộ trọng số w và hệ số b sao cho tại mỗi điểm dữ liệu trong tập X sẽ ánh xạ vào 1 điểm giá trị y tương ứng khớp với phân bố của dữ liệu.



3.3.2. Hồi quy bình phương tối thiểu từng phần (Partial Least Squares Regression)

Hồi quy bình phương tối thiểu từng phần là một kĩ thuật được sử dụng rộng rãi trong lĩnh vực hóa học, đặc biệt là trong trường hợp mà số lượng các biến độc lập lớn hơn nhiều so với số lượng dữ liệu. Đây là một phương pháp tối ưu hóa để lựa chọn một đường khớp nhất cho một dải dữ liệu ứng với cực trị của tổng các sai số thống kê giữa đường khớp và dữ liệu. Chính vì vậy, nhóm nhận thấy phương pháp này rất phù hợp với bài toán phổ NIR vì dữ liệu độc lập (X) có số đặc trưng khá lớn ứng với các dải bước sóng trong khi số mẫu thì hạn chế.

Phương pháp này giả định các sai số (error) của phép đo đạc dữ liệu phân phối ngẫu nhiên. Định lý Gauss-Markov chứng minh rằng kết quả thu được từ pháp pháp này không thiên vị và sai số của việc đo đạc dữ liệu không nhất thiết phải tuân theo.

3.3.3. Kết quả thực nghiệm

Quá trình thực nghiệm được tiến hành trên tập dữ liệu nêu trên sau khi nhóm dữ liệu theo 4 loại xoài: Cengkir (18 mẫu), Kent (85 mẫu), Kweni (29 mẫu) và Palmer (54 mẫu). Dữ liệu được xử lý dữ liệu trống và tiền xử lý làm mịn. Đối với dữ liệu mỗi loại, nhóm chia tập huấn luyện và kiểm thử với tỉ lệ 80:20.

***Bảng 2:** Bảng kết quả thực nghiệm dự đoán nồng độ các chất có trong quả trên các mô hình học máy.*

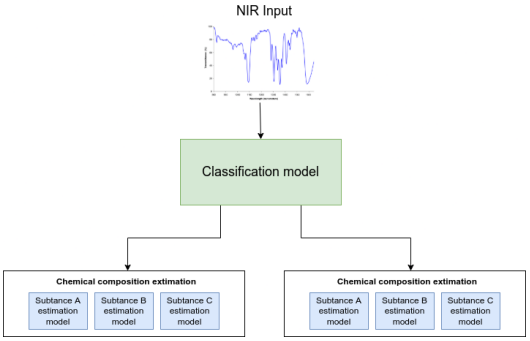
STT	Mô hình	Vit C	TA	SSC
1	Linear Regression	0.58	0.68	0.54
2	Partial Least Squares Regression	0.18	0.28	0.19

4. Bàn luận

4.1. Dự đoán thành phần hóa học trên nhiều chất khác nhau

Thử nghiệm trên cho thấy tính khả thi khi sử dụng phổ NIR để phân tích thành phần hóa học trong một loại thực phẩm nhất định. Tuy nhiên để ứng dụng vào thực tế, mô hình cần phải dự đoán được thành phần hóa học trên nhiều loại thực phẩm khác nhau. Cách thủ công nhất có thể thực hiện đó là xây dựng một mô hình phân loại từng loại thực phẩm, sau đó xây dựng nhiều mô hình học máy cho từng loại thực phẩm và từng loại chất cần dự đoán.

Hình x cho ta thấy ý tưởng ban đầu khi ta muốn dự đoán thành phần hóa học trong nhiều chất. Dễ dàng thấy khi thực hiện như ý tưởng trên, chúng ta cần xây dựng nhiều mô hình học máy. Đồng thời cần đi phân tích , trích xuất đặc trưng và tối ưu hóa cho nhiều mô hình học máy. Chúng ta cũng có thể thấy ở phương này không tận dụng được điểm tương đồng giữa các mô hình nếu có mối liên hệ với nhau.



Hình 16. Mô hình dự đoán thành phần hóa học đối với nhiều chất

4.2. Đề xuất phương pháp Multitask Learning và

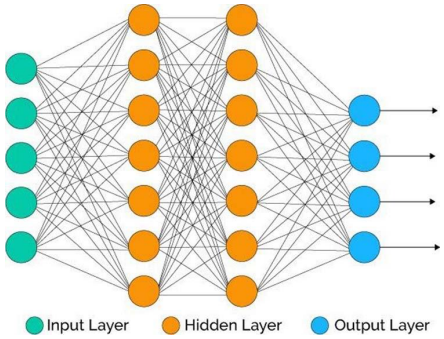
Neural Network

4.2.1. Neural Network

Neural Network (ANN) là mạng sử dụng các mô hình toán học phức tạp để tính toán thông tin. Trong ANN, các nút được liên kết với nhau và một mạng lưới các nút liên kết tạo thành một mạng Neural Network. Tương tự như bộ não người, ANN sử dụng một loạt các thuật toán phức tạo để xác định và nhận ra các mối quan hệ trong tập dữ liệu.

Kiến trúc một mạng ANN thường có 3 thành phần chính: Lớp đầu vào: Là lớp bên trái cùng, nơi dữ liệu được đưa vào mạng; Lớp đầu ra: Là lớp bên phải cùng, nơi nhận dữ liệu đầu ra; Lớp ẩn: Là những lớp ở giữa, thể hiện cho quá trình suy luận của mạng.

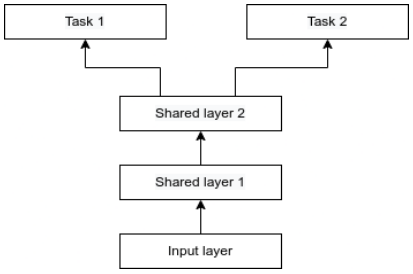
Trong ANN, quá trình lan truyền thuận giúp chứng ra đưa từ dữ liệu đầu vào, thông qua các trọng số là liên kết giữa các nút để nhận được đầu ra. Trong quá trình huấn luyện, đầu ra này được kết hợp với nhãn đúng để tạo ra hàm mất mát, từ đó tối ưu hóa các trọng số nhằm minimize giá trị hàm mất mát. Từ đó quá trình training sẽ giúp mô hình thông minh hơn.



Hình 17. Kiến trúc cơ bản mạng ANN

4.2.2. Multitask learning

Trong các bài toán học máy, chúng ta thường sử dụng chúng để giải quyết một vấn đề duy nhất. Điều này có thể làm chúng ta bỏ lỡ những thông tin có liên quan giữa các nhiệm vụ khác nhau. Đối với Multitask learning, chúng ta có thể chia sẻ cách trích xuất thông tin giữa các nhiệm vụ với nhau, nhờ vậy mô hình của chúng ta sẽ có khả năng tổng quát hóa cao hơn. Đồng thời, việc huấn luyện nhiều nhiệm vụ cùng lúc khiến mô hình hạn chế việc overfitting với một nhiệm vụ riêng lẻ.



Hình 18. Mô tả kiến trúc multitask learning

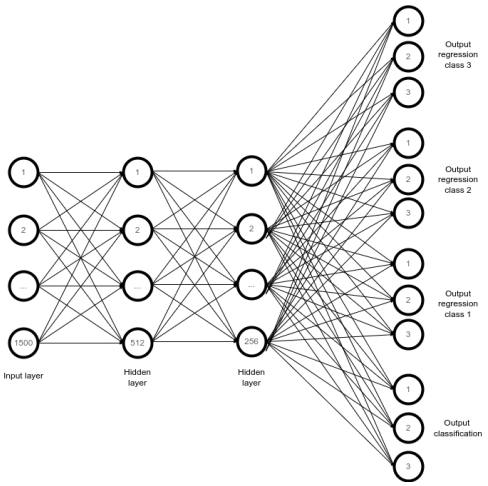
Khi xây dựng mô hình multitask learning đồng nghĩa với việc chúng ta đang làm việc với nhiều hàm mất mát, vì cậy cần có một hàm mất mát cuối cùng là tổng trọng số của các hàm mất mát của mỗi nhiệm vụ.

$$L_{final} = \sum_i \lambda_i L_i \quad (1)$$

Với L_{final} là hàm mất mát cuối cùng của mô hình. λ_i là trọng số của hàm mất mát L_i.

4.2.3. Ứng dụng Multitask Learning và Neural Network vào bài toán dự đoán nhiều thành phần hóa học nhiều chất

a. Kiến trúc mô hình



Hình 19. Kiến trúc mô hình đề xuất

Chúng tôi xây dựng kiến trúc mô hình gồm lớp input với 1500 nút tương ứng với 1500 khoảng bước sóng theo như dataset. 2 lớp ẩn tiếp theo để trích xuất đặc trưng phục vụ cho cả nhiệm vụ phân tích thành phần hóa học và phân loại thực phẩm. Tại 2 lớp này chúng tôi sử dụng hàm kích hoạt ReLU với công thức:

$$f(x) = \max(0, x) \quad (2)$$

Hình ảnh trên mô tả cho mô hình thực hiện hai nhiệm vụ: Phân loại loại thực phẩm, dự đoán thành phần hóa học cho từng loại thực phẩm.

Giả sử chúng ta cần phân tích n thành phần hóa học trong m loại thực phẩm. Số nút của lớp đầu ra là m + m*n. Trong đó m nút dùng để thực hiện nhiệm vụ phân loại và m*n nút còn lại thực hiện nhiệm vụ phân tích n thành phần của m loại.

Đối với m nút phân loại, chúng tôi sử dụng hàm mất mát Cross Entropy - là hàm mất mát sử dụng rất thông dụng trong nhiệm vụ phân loại. Hàm Cross Entropy có công thức:

$$L_{CE} = -\sum_{i=1}^{n} t_i log(p_i) \quad (3)$$

Với n là số lớp, t_i là nhãn của lớp i, p_i là xác suất dự đoán vào lớp i. Để có được xác xuất p_i. Ta cho kết quả tại lớp output qua hàm Softmax:

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_i}} \quad (4)$$

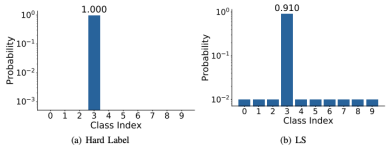
Với K là số lớp, z_i là kết quả của lớp output tại lớp i.

Vì tập dữ liệu sử dụng trong lần nghiên cứu này phân bố không đồng đều giữa các loại thực phẩm, chúng tôi có sử dụng thêm kỹ thuật class weight nhằm cân đối lại loss giữa các lớp. Khi đó hàm loss sẽ có trọng số giữa các lớp khác nhau với tác dụng phạt nặng hơn tại những lớp có ít

dữ liệu để cân bằng lại với những mẫu khác. Trọng số của loss từng lớp được tính theo công thức:

$$Class\ weight = 1 - \frac{Number\ of\ samples\ of\ the\ class}{Total\ number\ of\ samples} (5)$$

Ngoài ra, chúng tôi còn sử dụng thêm kỹ thuật Label Smoothing để tạo nhiễu cho nhãn. Với công thức hàm mất mát như trên, chúng ta có thể thấy hàm chỉ tối ưu hóa tại vị trí nhãn đúng với giá trị bằng 1, bỏ qua các nhãn sai với giá trị bằng 0.



Hình 20. So sánh nhãn thông thường và nhãn làm trơn

Kỹ thuật Label Smoothing thay vì sử dụng sử dụng nhãn 1-0 tại mỗi lớp (Hình bên trái), chúng sử dụng nhãn mới theo công thức

$$y2 = y1 * (1 - \varepsilon) + \varepsilon / n (6)$$

Với y2 là vector nhãn mới, y1 là vector nhãn ban đầu, ε là hệ số làm trơn và n là tổng số lớp.

Đối với mỗi n nút tương ứng với n thành phần hóa học của một chất. Chúng tôi sử dụng hàm mất mát Mean Absolute Error với công thức:

$$L_{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} (7)$$

Với n là số mẫu, y_i là giá trị dự đoán, x_i là giá trị nhãn.

Hàm mất mát này thường được dùng trong các bài toán dự đoán giá trị hồi quy tuy nhiên các giá trị có chứa những giá trị outlier.

Hàm loss cuối cùng của mô hình là tổng trọng số của 2 hàm mất trên với trọng số tương ứng là 0.5 và 0.5:

$$L = 0.2 * L_{CE} + 0.8 * L_{MAE}$$

b. Huấn luyện

Trong quá trình huấn luyện, chúng tôi thực hiện huấn luyện cả hai nhiệm vụ phân loại và dự đoán giá trị tuyến tính cùng một lúc. Trong đó nhiệm vụ tuyến tính sẽ đặc biệt hơn khi chỉ có n nút tương ứng với loài thực phẩm tương ứng được tối ưu hóa.

Chúng tôi chia việc huấn luyện thành hai giai đoạn.

Giải đoạn 1: Chúng tôi huấn luyện cả mô hình để mạng có khả năng trích xuất đặc trưng cho cả nhiệm vụ phân loại và phân tích thành phần.

Giai đoạn 2: Trong giai đoạn này, chúng tôi đóng băng các lớp trích xuất đặc trưng chúng, chỉ huấn luyện các lớp riêng của mỗi nhiệm vụ để tăng độ chính xác cho mỗi nhiệm vụ.

Quá trình huấn luyện được thực hiện trên cấu hình và thông số sau:

Gian đoạn 1:

Số epoch: 2000

Learning rate: 0.01

Gian đoạn 2:

Số epoch: 40000

Learning rate: 0.1 giảm dần trong quá trình huấn luyện.

c. Kết quả đánh giá

Chúng tôi sử dụng 2 metric chính để đánh giá:

- Để đánh giá quá trình phân loại: Sử dụng accuracy, có cách tính bằng tổng số dự đoán đúng chia cho tổng số dự đoán
- Để đánh giá quá trình phân tích thành phần hóa học: Sử dụng Mean absolute percentage error, có cách tính:

$$\text{MAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

Trong đó A_t là giá trị nhãn, F_t là giá trị dự đoán, n là số mẫu.

Bảng 3. Kết quả phân loại

Tên loại	Accuracy	Số lượng
Palmer	0.82	17
Cengkir	1	4
Kent	0.91	11
Kweni	1	6

Bảng 4. Đánh giá nhiệm vụ phân tích thành phần chính

	Vit C	**TA**	**SSC**
MAPE	0.24	0.40	0.24

Bảng 5. Kết hợp cả 2 tác vụ

	Vit C	**TA**	**SSC**
MAPE	0.26	0.39	0.26

5. Kết luận

Trong bài nghiên cứu này, chúng tôi đã nghiên cứu và thử nghiệm thành công mô hình có khả năng phân

tích thành phần hóa học của nhiều chất của nhiều loài thực phẩn dựa trên phổ NIR của loại thực phẩm đó. Kết quả cho thấy tính khả quan khi áp dụng dự án vào thực tế phục vụ cho việc đánh giá chất lượng thực phẩm. Tuy nhiên trong nghiên cứu này chúng tôi đang thử nghiệm trên bộ dữ liệu khá nhỏ đồng thời phân bố không đồng đều về số mẫu giữa các lớp cũng như phân bố không đồng đều theo thành phần hóa học, vì vậy dẫn đến hiệu suất của mô hình không cao đồng thời mô hình chưa mang tính tổng quát hóa cao. Trong tương lai, chúng tôi sẽ tự xây dựng bộ dữ liệu riêng về phổ NIR và thành phần hóa học nhằm tiếp tục nghiên cứu để đưa dự án tiến gần hơn với việc sử dụng trong thực tế.

Tài liệu tham khảo

[1] Agus Arip Munawar, Kusumiyati, Devi Wahyuni, Near infrared spectroscopic data for rapid and simultaneous prediction of quality attributes in intact mango fruits, 2019, doi: 10.1016\/j.dib.2019.104789.

[2] Dao Quang Huy, “Multi Task Learning - Một Số Điều Bạn Nên Biết,” Viblo, Oct. 19, 2020. https://viblo.asia/p/multi-task-learning-mot-so-dieu-ban-nen-biet-3P0lPD08lox (accessed Nov. 24, 2022).

[3] J. Brownlee, How to Choose Loss Functions When Training Deep Learning Neural Networks - MachineLearningMastery.com,” MachineLearningMastery.com, Jan. 29, 2019.

[4] Haritha Thilakarathne, “Handling Imbalanced Classes with Weighted Loss in PyTorch,” NaadiSpeaks, Jul. 31, 2021. https://naadispeaks.wordpress.com/2021/07/31/handling-imbalanced-classes-with-weighted-loss-in-pytorch/ (accessed Nov. 25, 2022).

[5] R. Müller, S. Kornblith, G. Google, and B. Toronto, “When Does Label Smoothing Help?” [Online]. Available: https://arxiv.org/pdf/1906.02629.pdf