ỨNG DỤNG PHƯƠNG PHÁP PHÂN TÍCH PCA VÀ LDA CHO CÁC THAM SỐ HANSEN ĐỂ DỰ ĐOÁN ĐỘ TAN CỦA BITUMEN TRONG CÁC LOẠI DUNG MÔI KHÁC NHAU

Nguyễn Tuệ Anh, Ngô Thanh An*

Trường Đại học Công nghiệp Thực phẩm TP.HCM

*Email: *ngothanhan@gmail.com*

Ngày nhận bài: 04/5/2022; Ngày chấp nhận đăng: 15/7/2022

TÓM TẮT

Phương pháp PCA (Principle Component Analysis) đã được sử dụng để tiền xử lý dữ liệu tham số độ tan Hansen của bitumen trong 48 loại dung môi khác nhau, nhằm mục đích loại bỏ các hiện tượng đa cộng tuyến giữa các biến số cũng như đảm bảo tính đồng nhất phương sai của dữ liệu trước khi tiến hành phân tích LDA (Linear Discriminant Analysis). Sau khi tiền xử lý, dữ liệu được tiến hành phân tích LDA để xác định mô hình dự đoán và phân loại nhằm phục vụ cho bài toán xác định độ tan của bitumen. Để đánh giá hiệu quả dự đoán của mô hình, cả hai phương pháp: phân chia dữ liệu ngẫu nhiên và xác thực chéo đã được sử dụng. Kết quả cho thấy, khi sử dụng phương pháp phân chia dữ liệu ngẫu nhiên (tỷ lệ 70:30), các đại lượng như độ chính xác, độ lặp và độ nhạy đều thay đổi giữa các lần thực thi chương trình, trong khi, đối với xác thực chéo, các đại lượng này không bị thay đổi. Khi xác thực chéo với tham số CV (số lần xác thực chéo) bằng 8, độ chính xác, độ lặp và độ nhạy của mô hình lần lượt là 75, 80,2 và 68,75%. Ngoài ra, kết quả phân tích LDA cho các nguồn dữ liệu thô (chưa qua tiền xử lý), dữ liệu đã được quy tâm và chuẩn hóa, và dữ liệu đã qua xử lý PCA khi xác thực chéo ở CV bằng 8 đều cho các kết quả đánh giá hiệu quả của mô hình đều giống nhau.

Từ khóa: PCA, LDA, tham số độ tan Hansen, độ tan, bitumen.

1. MỞ ĐẦU

Đánh giá độ tan của một chất tan trong các loại dung môi khác nhau luôn là một bài toán cấp thiết và quan trọng trong lĩnh vực công nghiệp hóa học, đặc biệt là các ngành liên quan đến dược phẩm, dung môi hữu cơ, sơn, v.v… [1-4]. Ở phạm vi công nghiệp, để lựa chọn được một loại dung môi phù hợp tương ứng với một chất tan xác định, các nhà nghiên cứu thường phải tiến hành một số lượng lớn thí nghiệm thử – sai, sau đó từ số liệu thu nhận được sẽ có các khuyến nghị sử dụng một loại dung môi thích hợp sao cho vừa giúp hòa tan tốt chất tan, nhưng đồng thời phải đáp ứng các tiêu chí về môi trường, về an toàn hóa chất, v.v…[5, 6]. Trong lĩnh vực học thuật, các nhà khoa học gần đây thường sử dụng các tham số độ tan Hansen để làm cơ sở cho việc đánh giá và lựa chọn dung môi thích hợp tương ứng với một loại chất tan nào đó [7-11]. Tuy nhiên, để có thể xác định được các tham số này, đòi hỏi phải tiến hành thực nghiệm trên rất nhiều mẫu dung môi khác nhau. Bên cạnh đó, việc lựa chọn một dung môi hoặc dự đoán độ tan của chất tan trong dung môi còn phải được căn cứ trên một bộ dữ liệu sẵn có. Dữ liệu sẵn có càng phong phú thì kết quả dự đoán sẽ càng chính xác. Về cơ bản, đầu tiên các nhà nghiên cứu phải xây dựng được một bộ cơ sở dữ liệu về tham số Hansen của chỉ riêng các dung môi. Sau đó, nếu có sẵn tham số Hansen của chất tan thì việc đánh giá độ tan của chúng trong dung môi thông qua đại lượng khoảng cách tương tự HSP sẽ trở nên dễ dàng hơn.

Trong số những chất tan có nhiều ứng dụng thực tế, bitumen là một đối tượng thường được các nhà sản xuất công nghiệp quan tâm. Bitumen được sản xuất từ quá trình chưng cất dầu thô, với thành phần phức tạp bao gồm nhiều loại hydrocacbon có kích thước phân tử, với độ phân cực khác nhau. Thành phần khối lượng chủ yếu của bitumen bao gồm 80-88% là cacbon, 8-11% là hydrogen và một lượng nhỏ kim loại như vanadi và niken [12]. Đây là một loại vật liệu nhựa nhiệt dẻo có chi phí thấp đã được sử dụng rộng rãi với vai trò là chất kết dính trong các ứng dụng thường thấy trong đời sống và công nghiệp, ví dụ như đường và công trình giao thông, vật liệu chống thấm và các ván ốp trong công nghiệp xây dựng [13]. Đặc tính chung của bitumen trong các ứng dụng công nghiệp kể trên đều có liên quan mật thiết đến tính lưu biến của nó. Do vậy, trong nhiều năm qua các nhà khoa học đã bắt đầu quan tâm đến việc sử dụng các loại bitumen được biến tính bằng polymer. Lý do chính của việc này là nhằm cải thiện các đặc tính lưu biến, và đồng thời làm cho nó ít nhạy cảm hơn với nhiệt độ. Độ cứng của chất kết dính cần phải được duy trì hợp lý ngay cả khi nhiệt độ bề mặt của công trình rất cao vào những tháng mùa hè, nhưng cũng phải có độ mềm dẻo hợp lý ở nhiệt độ thấp của mùa đông. Một lý do khác để biến tính với polyme là để tăng độ bền của chúng. Độ bền của bitume sẽ cải thiện rất nhiều nếu chọn được loại polymer thích hợp. Nhiều loại polyme khác nhau đã được thử nghiệm để làm chất điều chỉnh cho bitume, tuy nhiên, chỉ một số ít trong số đó mới có khả năng sử dụng trong thương mại [14, 15]. Cho đến nay, có rất ít công cụ để dự đoán khả năng tương thích giữa polymer và bitume, vì vậy việc phát triển bitumen biến tính bằng polymer mới phần lớn phải được thực hiện trên cơ sở "thử và sai". Việc hiểu rõ hơn về bản chất thực sự và các đặc tính hòa tan của bitumen trong các loại dung môi khác nhau do vậy sẽ là một nhu cầu và đòi hỏi cấp thiết cả trong lĩnh vực học thuật, cũng như công nghiệp. Tác giả Redelius vào năm 2004 đã công bố nghiên cứu về thuật toán để xác định thông số HSP của bitumen dựa vào số liệu độ tan của bitumen trong 48 dung môi khác nhau [16]. Trong khi đó, trong nghiên cứu [17], Rios và cộng sự đã sử dụng phần mềm Microsoft Excel để xác định thông số HSP. Phương pháp này dễ sử dụng và cho kết quả tương tự với phần mềm chuyên biệt là HSPiP hoặc với các phần mềm khác. Thật ra, về mặt bản chất, các công trình mà những nhà khoa học này đã tiến hành đều liên quan đến bài toán phân loại và dự đoán kết quả trong phân tích và khai thác dữ liệu, mà cụ thể là dữ liệu đa biến. Thực tế, có rất nhiều phương pháp phân loại dữ liệu có thể được áp dụng trong lĩnh vực hóa học như thế này, ví dụ như phương pháp phân tích thành phần chính (PCA), phương pháp phân tích phân biệt tuyến tính (LDA), phương pháp phân tích phân cụm (HCA), phương pháp K-means v.v…Tùy thuộc vào đặc tính của từng bộ dữ liệu, cũng như yêu cầu xử lý, người ta mới có thể lựa chọn được phương pháp thích hợp. Trong số các phương pháp phân tích dữ liệu được liệt kê ở trên, có hai phương pháp thường được sử dụng, đó là phương pháp PCA và LDA. Nếu như phương pháp PCA có mục đích là phân loại dữ liệu của các biến số độc lập mà không có thêm các biến phụ thuộc, thì LDA cũng có mục đích phân loại dữ liệu của các biến độc lập, nhưng bên cạnh đó còn có thêm các biến phụ thuộc đi kèm. Người ta thường gọi PCA là một phương pháp phân loại không giám sát, trong khi LDA là phương pháp phân loại có giám sát. Ngoài ra, PCA còn là một phương pháp rất hay được sử dụng để loại bỏ hiện tượng đa cộng tuyến (multicolinearity) thường xảy ra giữa các biến trong tập dữ liệu.

Xuất phát từ thực tế khảo sát độ tan của chất tan có thành phần phức tạp như bitumen, thì bộ dữ liệu luôn bao gồm các biến số độc lập chứa đựng các thông tin về năng lượng tương tác phân tán, năng lượng tương tác lưỡng cực và năng lượng tương tác hydro. Ngoài các biến độc lập như vậy, dữ liệu luôn đính kèm theo một biến phụ thuộc để mô tả về khả năng tan hoặc không tan của bitumen trong dung môi. Với những đặc trưng như thế, chỉ có phương pháp phân tích phân biệt LDA mới đáp ứng được khả năng phân loại dữ liệu. Tuy nhiên, không phải tất cả mọi dữ liệu đều có thể đưa vào để phân tích LDA. Trước khi phân tích, dữ liệu cần phải thỏa mãn một số giả thiết thống kê như dữ liệu phải có sự phân bố chuẩn (phân bố Gaussian), phải có sự đồng nhất phương sai (homoscedasticity), phải không có hiện tượng đa cộng tuyến,

v.v [18, 19]. Trong khuôn khổ nghiên cứu này, phương pháp PCA sẽ được sử dụng trước tiên để giúp tiền xử lý dữ liệu các tham số độ tan Hansen được trình bày trong cuốn sổ tay Hansen solubility parameters - A user's handbook [20]. Kế tiếp, các giá trị thu được sau PCA (các giá trị score) mới được đưa vào phân tích LDA.

2. DỮ LIỆU VÀ PHƯƠNG PHÁP

2.1. Dữ liệu tham số độ tan Hansen của các dung môi hữu cơ

Dữ liệu để phân tích LDA được tham khảo ở bảng 9.3 - trang 162 trong sổ tay Hansen [20].

Bảng 1. Tham số Hansen của một số dung môi được sử dụng để xác định độ tan của bitumen

STT	Dung môi	D	P	H	Khả năng tan
1	Benzophenone	19,6	8,6	5,7	1
2	2-Butanol	15,8	5,7	14,5	0
3	2-Butyl octanol	16,1	3,6	9,3	0
4	Caprolactone	19,7	15	7,4	0
5	Butyraldehyde	15,6	10,1	6,2	0
6	1-Chloro pentane	16	6,9	1,9	1
7	Chloroform	17,8	3,1	5,7	1
8	Cyclohexanol	17,4	4,1	13,5	0
9	Cyclohexanone	17,8	6,3	5,1	1
10	Cyclohexylamine	17,2	3,1	6,5	1
..	…	..	..	..	..

2.2. Các tham số độ tan Hansen

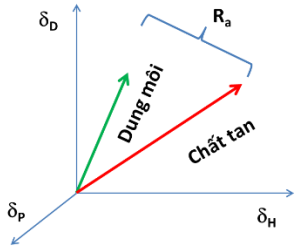
Lý thuyết về các tham số độ tan Hansen được công bố bởi Charles M.Hansen từ năm 1967 [21-23], với ý tưởng rằng chất tan và dung môi phải có cùng tính chất với nhau thì mới có khả năng tan lẫn vào nhau. Cùng tính chất với nhau được hiểu rằng một phân tử loại này liên kết với một phân tử loại khác nhưng lại có cùng kiểu liên kết như chính bản thân những phân tử cùng loại liên kết với nhau. Theo lý thuyết đề xuất bởi Hansen, khi xem xét độ tan của một chất tan trong dung môi, mỗi một phân tử bất kỳ luôn có một bộ tham số HSP đặc trưng của riêng mình, cụ thể bao gồm 3 đại lượng sau: (1) tham số tương tác phân tán (δ_D), (2) tham số tương tác lưỡng cực (δ_P), và (3) tham số tương tác hydro (δ_H). Về mặt hình thức, chúng ta có thể xem rằng mỗi một phân tử sẽ được biểu diễn bằng một điểm tọa độ trong không gian 3 chiều HSP (δ_D,δ_P,δ_H). Mỗi một tham số đều có thể tính được dựa trên các thông số liên quan đến các đặc trưng phân tử như nhiệt hóa hơi, chiết suất, moment lưỡng cực, hằng số điện môi [20].

Từ các tham số thành phần như vậy, có thể tính được tham số độ tan Hansen thông qua biểu thức:

$$\delta^2 = \delta_D^2 + \delta_P^2 + \delta_H^2$$

Nếu chỉ dừng lại ở tính toán các tham số độ tan Hansen thì chúng ta chỉ biết được vị trí tọa độ của phân tử trong không gian HSP mà thôi. Trong khi đó, quá trình hòa tan của một chất tan vào trong dung môi là một quá trình phức tạp mà ở đó phải xét đến sự tương tác giữa

chất tan với chính nó, giữa bản thân dung môi với nhau, và giữa dung môi với chất tan. Nguyên tắc cơ bản để cho rằng một chất tan sẽ tan trong dung môi, đó là chúng phải có cùng bản chất, hoặc chúng phải "tương tự" nhau. Để đặc trưng cho mức độ tương tự này, người ta thường sử dụng đại lượng khoảng cách HSP giữa hai phân tử, R_a hay còn gọi là bán kính tương tác. Hình 1 dưới đây mô tả về đại lượng R_a.



Hình 1. Khoảng cách tương tự HSP

R_a sẽ là thước đo mức độ giống nhau giữa hai phân tử. R_a càng nhỏ, hai phân tử (chất tan và dung môi) càng có nhiều khả năng tương thích với nhau. Công thức thông dụng thường được sử dụng để tính R_a chính là:

$$R_a^2 = 4\times\left(\delta_{D_1} - \delta_{D_2}\right)^2 + \left(\delta_{P_1} - \delta_{P_2}\right)^2 + \left(\delta_{H_1} - \delta_{H_2}\right)^2$$

Trong thực tế, để tính toán giá trị R_a không hề dễ dàng, bởi vì muốn xác định được một cách trực tiếp các tham số độ tan Hansen của một chất tan phức tạp, và bất kỳ là bất khả thi. Trong khi đó, các tham số Hansen của dung môi thường sẽ được xác định dễ dàng, và đầy đủ hơn, đồng thời những số liệu này sẽ được trình bày dưới dạng số liệu trong các bảng tra. Nhằm khắc phục tình trạng thực tế này, các nhà nghiên cứu phải tiến hành thực nghiệm để khảo sát độ tan của chất tan đang nghiên cứu trong các loại dung môi sẵn có (sẵn có, có nghĩa là đã xác định được đầy đủ các tham số Hansen thành phần). Với một chất tan cho trước, đầu tiên, người ta sẽ hòa tan vào dung môi sẵn có thứ nhất và sau đó quan sát hoặc đo lường độ tan của chất tan trong dung môi này. Nếu chất tan tan hoàn toàn trong dung môi này thì sẽ gán nhãn số 1 cho nó, còn nếu không tan thì dung môi sẽ được gán nhãn 0. Quá trình này được lặp lại tùy theo số lượng dung môi sẵn có, sau đó ghi nhận số liệu tan hoặc không tan (tương ứng với nhãn 1 hoặc 0). Sau khi hoàn thành các khảo sát này, chúng ta sẽ xây dựng được một mặt cong không gian đi qua tọa độ của các dung môi có khả năng hòa tan tốt chất tan. Một khi đã có đầy đủ thông tin về mặt cong không gian này thì nếu có một dung môi bất kỳ nào đó với bộ tham số Hansen có sẵn, người ta có thể dựa vào vị trí tọa độ của dung môi đó trong không gian HSP vừa xây dựng ở trên để đánh giá xem liệu rằng nó có phải là dung môi thích hợp để hòa tan chất tan hay không. Ở đây, cần phải nhấn mạnh rằng mặt cong mà chúng ta thu được chỉ tương ứng với loại chất tan mà ta đang khảo sát. Mặt cong này chỉ đi qua các điểm tọa độ của những dung môi nào tan tốt chất tan mà thôi. Nếu điểm tọa độ của dung môi đang xem xét nằm trên mặt cong thu được ở trên thì dung môi đó sẽ hòa tan tốt chất tan. Điều này có thể lí giải dựa vào việc định nghĩa sự tương đồng giữa dung môi và chất tan.

2.3. Phương pháp PCA (Principal Component Analysis)

Trong số các kỹ thuật phân tích đa biến, PCA được sử dụng thường xuyên nhất vì nó là điểm khởi đầu trong quá trình khai thác dữ liệu. Nó nhằm mục đích giảm thiểu kích thước của dữ liệu, giúp thực hiện việc phân loại dữ liệu, cũng như giúp tìm kiếm các mối tương quan (nếu có) giữa các biến số. PCA có thể được coi là phương pháp đi tìm một hệ cơ sở trực chuẩn đóng vai trò một phép xoay, sao cho trong hệ cơ sở mới này, phương sai theo một số chiều nào đó là rất nhỏ, và ta có thể bỏ qua, ta chỉ cần giữ lại các chiều/thành phần khác quan trọng

hơn. Một ứng dụng rất hữu ích khác của PCA nữa, đó là giúp giảm hiện tượng đa cộng tuyến giữa các biến [24]. Hiện tượng đa cộng tuyến là hiện tượng thường gặp trong các phép phân tích hồi quy, khi mà các biến độc lập có mối tương quan rất mạnh với nhau. Nếu một mô hình hồi quy xảy ra hiện tượng đa cộng tuyến sẽ làm cho nhiều chỉ số bị sai lệch, dẫn đến kết quả của việc phân tích định lượng không còn mang lại nhiều ý nghĩa [25].

2.4. Phương pháp LDA (Linear Discriminant Analysis)

Việc xác định một sản phẩm hoặc một chất hóa học nào đó trong phạm vi mong đợi thường được thực hiện bằng cách phân tách các dữ liệu thành các lớp khác nhau. Các lớp khác nhau được phân loại dựa vào tên nhãn của nó. Dữ liệu để phân tích LDA trong nghiên cứu này chính là khả năng tan của bitumen với hai nhãn 1 (tan) và 0 (không tan) trong 48 loại dung môi khác nhau. Trong Bảng 1 được trình bày ở trên, thì các tham số Hansen là các biến độc lập mô tả các đặc tính của bitumen về tương tác phân tán (D), tương tác lưỡng cực (P) và tương tác hydro (H), trong khi khả năng tan là biến phụ thuộc, chính là nhãn được gán cho nhóm.

Phương pháp LDA sử dụng các nhãn để giảm kích thước, đồng thời được thiết kế để tối đa hóa khoảng cách giữa các lớp [18, 19]. Về mặt nguyên tắc, để thực hiện LDA, khoảng cách giữa các nhóm dữ liệu sẽ được tối đa hóa thông qua việc tối đa khoảng cách giữa hai kỳ vọng của hai nhóm và đồng thời tối thiểu độ lệch trong nội bộ của nhóm dữ liệu.

Về mặt nguyên tắc, nhiệm vụ của phân tích LDA nói một cách đơn giản đó là đi tìm vector chiếu w sao cho giá trị J(w) đạt cực đại, với J(x) được phát biểu thông qua phương trình dưới đây:

$$J(x) = \frac{(\tilde{\mu}_1 - \tilde{\mu}_2)^2}{(\tilde{s}_1 + \tilde{s}_2)} = \frac{d^2}{(\tilde{s}_1 + \tilde{s}_2)} \qquad (1)$$

Trong đó, w: là vector chiếu được sử dụng để chiếu x lên y. $\tilde{\mu}_1$ và $\tilde{\mu}_2$ là kỳ vọng trong nhóm dữ liệu 1 và 2. $\tilde{s}_1$ và $\tilde{s}_2$ là độ lệch trong mỗi nhóm, d là khoảng cách giữa các lớp

2.5. Trình tự thực hiện phân tích LDA

Việc phân tích LDA được thực hiện thông qua sử dụng ngôn ngữ lập trình python – version 3.9.6 dành cho Windows.

Trình tự phân tích bao gồm các bước sau:

- Nhập dữ liệu
- Thiết lập tham số cho phân tích LDA
- Đánh giá mô hình thu được

2.6. Đánh giá mô hình

Để đánh giá hiệu quả của mô hình phân loại dữ liệu thu được, hiện nay, có nhiều phương pháp đã được sử dụng như phương pháp phân chia dữ liệu ngẫu nhiên, hoặc phương pháp xác thực chéo.

Phương pháp phân chia dữ liệu ngẫu nhiên được tiến hành bằng cách phân chia bộ dữ liệu một cách ngẫu nhiên theo một tỷ lệ phần trăm xác định. Trong khi phần dữ liệu được phân tách với tỷ lệ lớn sẽ được sử dụng để huấn luyện và xây dựng mô hình, thì phần nhỏ còn lại sẽ được dùng để kiểm tra hiệu quả của mô hình. Đối với kích thước dữ liệu phù hợp, quá trình phân chia dữ liệu huấn luyện và kiểm tra được thực hiện theo tỷ lệ 70:30. Điều đó có nghĩa là sẽ dành 70% dữ liệu cho huấn luyện, còn lại 30% dành cho kiểm tra mô hình. Đây là một tỷ

lệ rất thường được sử dụng trong phân tích dữ liệu, do vậy, trong nghiên cứu này, tỷ lệ 70:30 sẽ được lựa chọn để phân chia dữ liệu phục vụ cho quá trình phân tích LDA.

Mặt khác, phương pháp xác thực chéo k lần được bắt đầu bằng cách phân tách ngay từ đầu hai bộ dữ liệu trong đó một bộ sẽ dành cho thử nghiệm, bộ còn lại sẽ được dùng cho việc xác thực. Bộ dữ liệu dành cho thử nghiệm lại được phân tách thành k đoạn nhỏ, sau đó, một đoạn nhỏ lại được dành riêng ra cho việc kiểm tra cho (k-1) đoạn còn lại. Quá trình sẽ được lặp lại cho đến khi tất cả k đoạn nhỏ đó đã được dùng để kiểm tra cho phần dữ liệu còn lại. Có thể thấy rằng trong xác thực chéo, tập dữ liệu sẽ không được phân chia thành tập dữ liệu huấn luyện và thử nghiệm chỉ bởi một lần duy nhất. Thay vào đó, người ta sẽ liên tục phân vùng tập dữ liệu thành các nhóm nhỏ hơn và sau đó tính trung bình hiệu suất trong mỗi nhóm.

3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Kiểm định các giả thiết thống kê để thực hiện LDA

Để tiến hành phân tích LDA, dữ liệu cần phải thỏa mãn một số giả thiết sau: thứ nhất, các biến số phải tuân theo phân bố chuẩn; thứ hai, các biến số phải có phương sai đồng nhất; và thứ ba, không tồn tại các hiện tượng đa cộng tuyến giữa các biến [18, 19].

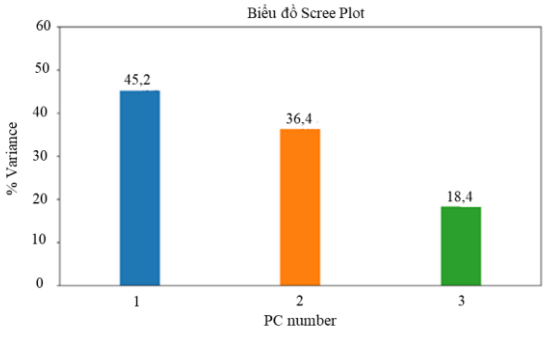
Bảng 2. Các thông số kiểm định thống kê cho dữ liệu ban đầu (α = 0,01)

	p-value (phân bố Gaussian) cho từng biến	VIF	p-value (kiểm định Levene) cho cả ba biến số
Biến phân tán, D	0,149	4,52	0,0012
Biến tương tác lưỡng cực, P	0,199	3,81	
Biến tương tác hydro, H	0,355	3,83	

Các giá trị p-value đều lớn hơn giá trị α, như vậy có thể thấy rằng dữ liệu này tuân theo phân bố chuẩn. Các giá trị VIF (Variance Inflation Factor) của các biến số đều lớn hơn 2, điều này cho thấy có sự tồn tại của hiện tượng đa cộng biến trong dữ liệu [26]. Giá trị p-value của kiểm định Levene nhỏ hơn cả giá trị α, cho thấy rằng tính đồng nhất phương sai của dữ liệu là không thỏa mãn.

3.2. Phân tích PCA

Biểu đồ Scree từ Hình 2 dưới đây cho thấy độ tích lũy của 3 thành phần chính là 100%. Mặt khác, dữ liệu ban đầu cũng chỉ có 3 đặc trưng (P, D và H). Do vậy, sau khi tiến hành phân tích PCA, sẽ giữ lại 3 thành phần chính.



Hình 2. Biểu đồ Scree khi phân tích PCA cho các tham số Hansen

Sau khi tiền xử lý bằng PCA, dữ liệu score thu được sẽ được dùng để đánh giá lại các thông số kiểm định thống kê nhằm kiểm tra sự đáp ứng các giả thiết cho quá trình phân tích LDA.

Bảng 3. Các thông số kiểm định thống kê cho dữ liệu sau khi được phân tích PCA (α = 0,01)

	p-value (phân bố Gaussian) cho từng biến	VIF	p-value (kiểm định Levene) cho cả ba biến số
Biến phân tán, D	0,913	1	0,0183
Biến tương tác lưỡng cực, P	0,019	1	
Biến tương tác hydro, H	0,496	1	

Số liệu từ Bảng 3 cho thấy rằng các giá trị p-value của cả phân bố Gaussian và kiểm định Levene đều lớn hơn giá trị α, điều đó có nghĩa là dữ liệu tham số Hansen sau xử lý PCA tuân theo phân bố chuẩn và có sự đồng nhất phương sai ở cả ba biến số D, P và H. Giá trị VIF thu được nhỏ hơn 2 là dấu hiệu cho thấy không còn tồn tại hiệu ứng đa cộng tuyến xảy ra giữa các biến sau xử lý PCA.

Như vậy, có thể thấy rằng, sau khi đã qua xử lý PCA, dữ liệu đã đáp ứng được các giả thiết về phân bố Gaussian, đồng nhất phương sai và không có hiện tượng đa cộng tuyến. Dữ liệu này hoàn toàn phù hợp để có thể dùng cho phân tích LDA.

3.3. Phân tích LDA

Dữ liệu đầu ra sau khi phân tích PCA cho các tham số Hansen sẽ được đưa vào phân tích LDA. Số phần tử (n) sử dụng để khai báo tham số đầu vào của LDA phải tuân thủ điều kiện:

n = min[số biến số độc lập, (số nhãn trong biến phụ thuộc) - 1]

Trong nghiên cứu này, chỉ có 3 biến độc lập D, P, H và 2 nhãn (0 và 1) của biến phụ thuộc, do vậy, số phần tử n sẽ bằng 1.

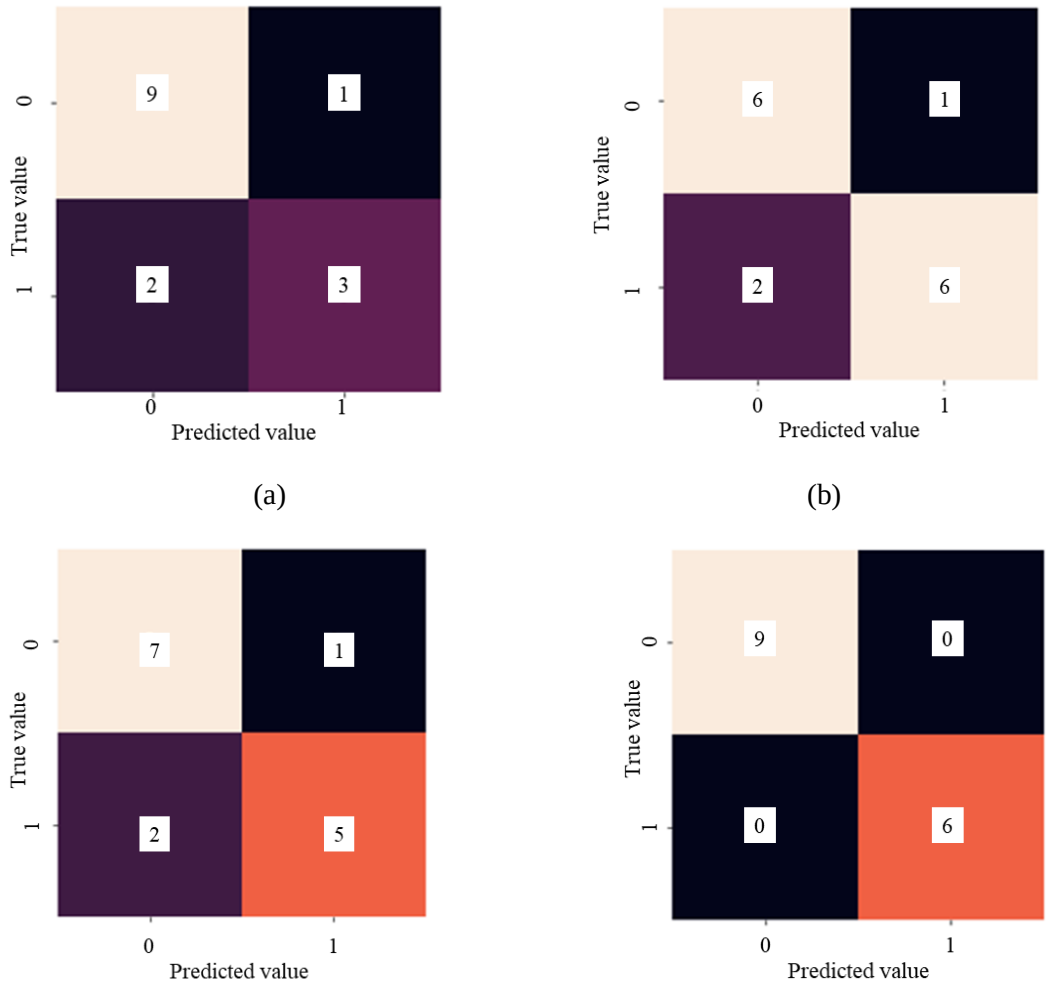
Mô hình phân loại dữ liệu thu được sau khi phân tích LDA sẽ được đánh giá thông qua hai phương pháp, bao gồm phương pháp phân chia dữ liệu ngẫu nhiên và phương pháp xác thực chéo.

3.3.1. Đánh giá mô hình thông qua phương pháp phân chia dữ liệu

Sau khi thực hiện phân tích LDA, ma trận lỗi sẽ được sử dụng để đánh giá chất lượng của quá trình phân loại dữ liệu. Bên cạnh đó, các đại lượng độ chính xác, độ lặp, và recall cũng được sử dụng để đánh giá đầy đủ hơn về mức độ hiệu quả của mô hình dự đoán LDA. Ma trận lỗi thể hiện có bao nhiêu điểm dữ liệu *thực sự* thuộc vào một lớp, và được dự đoán rơi vào một lớp. Thực ra, ma trận lỗi là một bảng đặc biệt (ma trận vuông) được dùng để minh họa hiệu quả của các thuật toán trong các bài toán phân loại.

Hình 3 là ma trận lỗi thu được qua 4 lần thực thi chương trình. Dựa vào kết quả ở Hình 3 (a), chúng ta sẽ thấy ở hàng một và cột một, với giá trị nhãn (true value) là 0, mô hình dự đoán (predicted value) là 0 với số lượng là 9 trong tổng số 10 mẫu (tức là đạt tỷ lệ đúng 90% cho nhãn 0). Tại hàng một và cột hai, có 1 mẫu bị dự đoán sai. Tại hàng hai và cột hai, mô hình dự đoán đúng 3 trong tổng số 5 mẫu, có nghĩa là đúng 60%. Tại hàng hai và cột một, nhãn có giá trị 1 bị dự đoán sai với số lượng là 2 trong tổng số 5 mẫu.

Như vậy có thể thấy giá trị tại hàng thứ i, cột thứ j là số lượng điểm dữ liệu lẽ ra phải thuộc vào lớp i nhưng lại được dự đoán là thuộc vào lớp j**.** Các phần tử trên đường chéo của ma trận là số điểm được phân loại đúng của mỗi lớp dữ liệu. Một mô hình tốt sẽ cho một ma trận lỗi có các phần tử trên đường chéo chính có giá trị lớn, các phần tử còn lại phải có giá trị nhỏ.



Hình 3. Ma trận lỗi thu được khi thực thi chương trình lần 1 (a), 2 (b), 3 (c) và 4 (d)

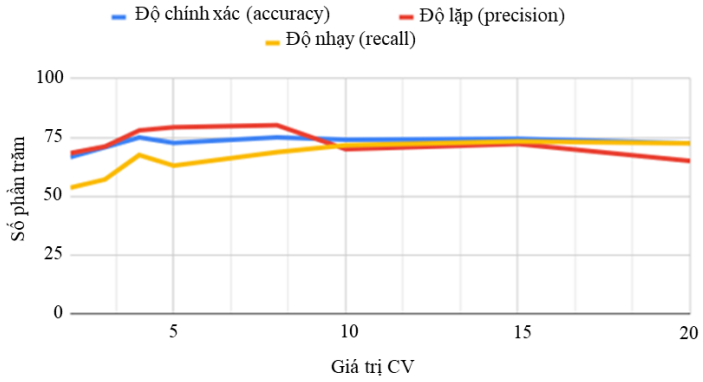
Kết quả từ Hình 3 cho thấy rằng các ma trận lỗi là hoàn toàn khác nhau cho mỗi lần thực hiện việc phân tích. Sở dĩ có điều này là bởi vì tập dữ liệu đã được phân chia một cách ngẫu nhiên thành hai tập con bao gồm: tập dữ liệu dành cho huấn luyện mô hình (tỷ lệ 70%) và tập dữ liệu dành cho kiểm tra mô hình (30%). Chính vì phân chia ngẫu nhiên như vậy, nên kết quả kiểm tra tính chính xác của thuật toán thông qua ma trận lỗi sẽ không bị trùng lặp. Thông qua ma trận lỗi, độ chính xác, độ lặp và độ nhạy của mô hình sẽ hoàn toàn xác định. Các thông số này được trình bày ở Bảng 4 dưới đây.

Bảng 4. Các thông số đánh giá hiệu quả mô hình

Lần thực hiện chương trình	Độ chính xác (accuracy)	Độ lặp (precision)	Độ nhạy (recall)
1	0,8	0,75	0,6
2	0,8	0,86	0,75
3	0,8	0,83	0,71
4	1,0	1,0	1,0

Kết quả ở Bảng 4 cho thấy sự không ổn định của các thông số đánh giá hiệu quả mô hình khi thực hiện việc phân chia dữ liệu ngẫu nhiên. Đây chính là một nhược điểm rất lớn của phương pháp kiểm tra mô hình này. Nhược điểm này còn trở nên nghiêm trọng hơn khi tập dữ liệu nhỏ, từ đó dẫn đến phương sai sẽ cao. Ngoài ra, do phân chia ngẫu nhiên, kết quả có thể hoàn toàn khác nhau đối với các lần thử nghiệm khác nhau. Điều này xảy ra là do trong một số phạm vi nhất định, các mẫu dễ phân loại sẽ được đưa vào tập kiểm tra, trong khi ở những phân vùng khác, tập kiểm tra lại được nhận những mẫu khó phân loại.

3.3.2. Đánh giá mô hình thông qua phương pháp xác thực chéo



Hình 4. Sự phụ thuộc của các thông số đánh giá hiệu quả mô hình đối với tham số CV

Kết quả xác thực chéo khi sử dụng các tham số CV (số lần xác thực chéo) thay đổi từ 0 đến 20 được trình bày ở Hình 4. Trong khoảng CV từ 5-10, có thể thấy rằng các đại lượng như độ chính xác, độ nhạy và độ lặp hầu như đạt giá trị cao nhất. Do vậy, đại lượng CV = 8 sẽ thích hợp cho phân tích LDA với dữ liệu loại này.

Bảng 5. Các thông số đánh giá hiệu quả mô hình khi sử dụng CV = 8

Lần thực hiện chương trình	Độ chính xác (accuracy)	Độ lặp (precision)	Độ nhạy (recall)
1	75	80,2	68,75
2	75	80,2	68,75
3	75	80,2	68,75

Bảng 5 thể hiện kết quả phân tích LDA với CV = 8 trong 3 lần khác nhau cho thấy các đại lượng dùng để đánh giá hiệu quả mô hình đều không đổi. Đối chiếu với kết quả thu được khi phân chia dữ liệu ngẫu nhiên, trong xác thực chéo, chúng ta có thể thấy rằng tập dữ liệu sẽ không được phân chia thành tập dữ liệu huấn luyện và thử nghiệm chỉ bởi một lần duy nhất. Thay vào đó, người ta sẽ liên tục phân vùng tập dữ liệu thành các nhóm nhỏ hơn và sau đó tính trung bình hiệu suất trong mỗi nhóm. Bằng cách này, sẽ giúp giảm tác động của tính ngẫu nhiên của phân vùng lên kết quả.

3.3.3. So sánh kết quả phân tích LDA đối với dữ liệu ban đầu và dữ liệu đã qua xử lý PCA

Cả ba nguồn dữ liệu, bao gồm dữ liệu ban đầu, dữ liệu ban đầu được quy tâm và chuẩn hóa, và dữ liệu sau xử lý PCA đều được đưa vào phân tích LDA, kết hợp với phương pháp xác thực chéo có CV = 8 như ở trên.

Bảng 6. Kết quả phân tích LDA cho các dữ liệu đầu vào khác nhau

Loại dữ liệu	Độ chính xác (accuracy)	Độ lặp (precision)	Độ nhạy (recall)
Dữ liệu ban đầu	75	80,2	68,75
Dữ liệu ban đầu được quy tâm và chuẩn hóa	75	80,2	68,75
Dữ liệu đã qua xử lý PCA	75	80,2	68,75

Qua Bảng 6, có thể nhận thấy rõ ràng rằng, mặc dù dữ liệu thô ban đầu có sự vi phạm các giả thiết thống kê để tiến hành LDA, nhưng kết quả thu được từ phân tích LDA cho dữ liệu thô và dữ liệu đã qua xử lý PCA đều cho các kết quả như nhau. Bên cạnh đó, việc tiền xử lý dữ liệu bằng phương pháp quy tâm và chuẩn hóa dữ liệu cũng không làm biến đổi các đại lượng độ chính xác, độ lặp và độ nhạy của mô hình. Điều này có thể được lí giải rằng phương pháp LDA không bị ảnh hưởng quá nhiều nếu dữ liệu ban đầu không đáp ứng được một số giả thiết như đã liệt kê trong phần 3.1 [27]. Hơn nữa, khi $1 < VIF < 5$ thì hiện tượng đa cộng tuyến là thấp và có thể chấp nhận được [26].

4. KẾT LUẬN

Phương pháp tiền xử lý dữ liệu bằng PCA đã giúp dữ liệu đầu vào của LDA đáp ứng các giả thiết thống kê bao gồm phân bố chuẩn, đồng nhất phương sai và hiện tượng đa cộng tuyến. Kết quả phân tích LDA với tham số độ tan Hansen của các dung môi khác nhau, kết hợp với dữ liệu khả năng tan của bitumen trong chính các dung môi này đã cho thấy khả năng dự đoán của mô hình là chấp nhận được. Phương pháp đánh giá hiệu quả của mô hình bằng cách phân chia ngẫu nhiên dữ liệu theo tỷ lệ 70:30 không có tính ổn định và khó có thể xử dụng, trong khi đó, phương pháp xác thực chéo đem đến các giá trị về độ chính xác, độ lặp và độ nhạy khá ổn định khi thực thi chương trình và có các giá trị lần lượt là 75, 80,2 và 68,75%.

TÀI LIỆU THAM KHẢO

1. Blumenroth D., Zumbühl S., Scherrer C., Müller W. - Sensitivity of modern oil paints to solvents. Effects on synthetic organic pigments. In: Issues in contemporary oil paint, Cham: Springer (2014) 351–362.
2. La Nasa J., Lee J., Degano I., Burnstock A., van den Berg K.J., Ormsby B., Bonaduce I. - The role of the polymeric network in the water sensitivity of modern oil paints, Scientific Reports **9** (2019) 1-12.
3. Banti D., La Nasa J., Tenorio L., Modugno F., van den Berg J., Lee J., Ormsby B., Burnstock A., Bonaduce I. - A molecular study of modern oil paintings: investigating the role of dicarboxylic acids in the water sensitivity of modern oil paints, RSC Adv. **8** (2018) 6001-6012.
4. Hancock B.C., Peter York P., Rowe R.C. - The use of solubility parameters in pharmaceutical dosage form design, Int. J. Pharm. **148** (1997) 1-21.
5. Lee J., Park S.A., Ryu S.U., Chung D., Park T., Son S.Y. - Green-solvent-processable organic semiconductors and future directions for advanced organic electronics, Journal of Materials Chemistry A **8** (2020) 21455-21473.

6. Cunningham M.F., Campbell J.D., Fu Z., Bohling J., Leroux J.G., Mabee W., Robert T. - Future green chemistry and sustainability needs in polymeric coatings, Green Chemistry **21** (2019) 4919-4926.
7. Guenthner A.J., Lamison K.R., Lubin L.M., Haddad T.S., Mabry J.M. - Hansen solubility parameters for octahedral oligomeric silsesquioxanes, Industrial & Engineering Chemistry Research **51** (2012) 12282-12293.
8. Batista M.M., Reginaldo Guirardello R., Krähenbühl M.A. - Determination of the Hansen solubility parameters of vegetable oils, biodiesel, diesel, and biodiesel-diesel blends, J. Am. Oil Chem. Soc. **92** (2015) 95-109.
9. Negera D., Yohannes T. - Hansen solubility parameters and green solvents for organic photovoltaics, Int. J. Adv. Sci. Res. Eng. **4** (2018) 128-129.
10. Benazzouz A., Moity L., Pierlot C., Sergent M., Molinier V., Aubry M. J. - Selection of a greener set of solvents evenly spread in the Hansen space by space-filling design, Industrial & Engineering Chemistry Research **52** (2013) 16585-16597.
11. Park W. J., Kim Y. M., Im I. S., Go S. K., Nho S. N., Lee B. K. - Development of correlations between deasphalted oil yield and Hansen solubility parameters of heavy oil SARA fractions for solvent deasphalting extraction, Journal of Industrial and Engineering Chemistry **107** (2022) 456-465.
12. Porto M., Caputo P., Loise V., Eskandarsefat S., Teltayev B., & Oliviero Rossi, C. Bitumen and bitumen modification: A review on latest advances, Applied Sciences **9** (2019) 742-776.
13. Redelius P., Soenen H. - Relation between bitumen chemistry and performance, Fuel **140** (2015) 34-43.
14. Zhu J., Birgisson B., Kringos N. - Polymer modification of bitumen: Advances and challenges. European Polymer Journal **54** (2014) 18-38.
15. Navarro F. J., Partal P., García-Morales M., Martín-Alfonso M. J., Martinez-Boza F., Gallegos C., Diogo A. C. - Bitumen modification with reactive and non-reactive (virgin and recycled) polymers: a comparative analysis, Journal of Industrial and Engineering Chemistry **15** (2009) 458-464.
16. Redelius, P. - Bitumen solubility model using Hansen solubility parameter, Energy & Fuels **18** (2004) 1087-1092.
17. Díaz de los Ríos M., Hernández Ramos E. - Determination of the Hansen solubility parameters and the Hansen sphere radius with the aid of the solver add-in of Microsoft Excel, SN Applied Sciences **2** (2020) 1-7.
18. Xanthopoulos P., Pardalos P.M., Trafalis T.B. - Linear discriminant analysis, In Robust data mining **5** (2013) 27-33.
19. Tharwat A., Gaber T., Ibrahim A., Hassanien A.E. - Linear discriminant analysis: A detailed tutorial, AI Communications **30** (2017) 169-190.
20. Hansen C.M. - Hansen solubility parameters - A user's handbook, CRC Press, 2nd Ed, (2007).
21. Hansen C.M. - The three dimensional solubility parameter - key to paint component af finities I, J. Paint Technol. **39** (1967) 104-117.
22. Hansen C.M. - The three dimensional solubility parameter - key to paint component af finities II, J. Paint Technol. **39** (1967) 505-510.

23. Hansen C.M., Skaarup K. - The three dimensional solubility parameter - key to paint component affinities III, J. Paint Technol. **39** (1967) 511-514.
24. Lafi S. Q., Kaneene J. B. - An explanation of the use of principal-components analysis to detect and correct for multicollinearity, Preventive Veterinary Medicine **13** (1992) 261-275.
25. Næs T., Mevik B. H. - Understanding the collinearity problem in regression and discriminant analysis, Journal of Chemometrics: A Journal of the Chemometrics Society **15** (2001) 413-426.
26. Shrestha N. - Detecting multicollinearity in regression analysis. American Journal of Applied Mathematics and Statistics **8** (2020) 39-42.
27. Büyüköztürk Ş., Çokluk-Bökeoğlu Ö. - Discriminant function analysis: Concept and application, Egitim Arastirmalari - Eurasian Journal of Educational Research **33** (2008) 73-92.

ABSTRACT

APPLICATION OF PCA AND LDA METHODS FOR HANSEN PARAMETERS IN PREDICTION OF BITUMEN SOLUBILITY IN DIFFERENT SOLVENTS

Nguyen Tue Anh, Ngo Thanh An*
Ho Chi Minh City University of Food Industry
*Email: *ngothanhan@gmail.com*

PCA method (Principle Component Analysis) was used to preprocess Hansen solubility parameters of bitumen in 48 different solvents, aiming to eliminate multicollinearities between variables as well as to ensure the homoscedasticity of the data. After preprocessing, the data were analyzed by LDA (Linear Discriminant Analysis) to determine a classification model for the predicting solubility of bitumen. Both methods: random split data and cross-validation were used to evaluate the predictive efficiency of the model. The results showed that, when using the random split data method (ratio 70:30), quantities such as accuracy, precision and recall were changed between program executions, while, in the case of cross-validation, these quantities were not. When cross-validating with the CV parameter (number of cross-validations) equaled to 8, the model's accuracy, precision and recall of the model were 75, 80.2 and 68.75%, respectively. In addition, the results of LDA analysis for raw data sources, centered and normalized data, and PCA-processed data, when cross-validated at the CV of 8, presented that the evaluation performance of the model was all the same.

Keywords: PCA, LDA, Hansen parameters, solubility, bitumen.