**Tóm tắt**

Dữ liệu từ mảng cảm biến khí (GSA) thường có đặc điểm là số chiều cao và số lượng mẫu nhỏ. Khi sử dụng bộ phân loại trực tiếp cho dữ liệu GSA, mô hình dễ bị quá khớp và tốn nhiều thời gian tính toán. Giải pháp truyền thống là giảm chiều dữ liệu trước khi phân loại. Tuy nhiên, việc chọn phương pháp giảm chiều phù hợp rất tốn thời gian và công sức, đồng thời một số đặc trưng hữu ích có thể bị mất sau khi giảm chiều — đặc biệt là đối với dữ liệu cảm biến phản hồi yếu với khí nồng độ thấp.

Trong bài báo này, chúng tôi đề xuất một khung học máy cực đoan dựa trên tập hợp đặc trưng (FE-ELM) để phân loại dữ liệu GSA. Trong FE-ELM, đầu tiên dữ liệu thời gian từ từng cảm biến được giảm mẫu, sau đó kết hợp các đặc trưng đã giảm mẫu từ nhiều cảm biến để tạo thành các tập hợp đặc trưng hợp nhất. Tiếp theo, một mô hình ELM cơ sở được huấn luyện độc lập trên từng tập hợp đặc trưng hợp nhất bằng cách giải bài toán bình phương tối tiểu. Dự đoán cuối cùng được đưa ra bằng cách bỏ phiếu kết quả dự đoán từ tất cả các mô hình cơ sở.

So với các phương pháp truyền thống, phương pháp này giải quyết được vấn đề quá khớp và có thể áp dụng trực tiếp cho dữ liệu GSA mà không cần giảm chiều đặc trưng trước. Hơn nữa, việc tập hợp các mô hình cơ sở mà không làm mất nhiều thông tin ban đầu giúp FE-ELM đạt hiệu suất phân loại hiệu quả và mạnh mẽ hơn. Kết quả thực nghiệm trên cả dữ liệu GSA tự thu thập với khí nồng độ rất thấp (ppb) và dữ liệu công khai xác nhận rằng phương pháp FE-ELM vượt trội hơn các phương pháp truyền thống và mở rộng giới hạn phát hiện của mảng cảm biến.

**I. GIỚI THIỆU**

Mảng cảm biến khí (Gas Sensor Array - GSA) là một thiết bị quan trọng dùng để phát hiện khí, chuyển đổi thông tin mùi thành tín hiệu điện, và cho phép thực hiện nhận dạng mẫu sau khi xử lý tín hiệu [1], [2]. Mỗi mẫu dữ liệu GSA bao gồm các chuỗi thời gian đa biến, thường có số chiều rất lớn (ít nhất hàng nghìn) [3]. Đồng thời, số lượng mẫu lại thường nhỏ (vài chục đến vài trăm) do thiếu quy trình tiêu chuẩn thu thập mẫu và chi phí cao [4].

Các bộ phân loại cổ điển được sử dụng trong phân loại dữ liệu GSA bao gồm k-nearest neighbors (KNN), máy vector hỗ trợ (SVM), rừng ngẫu nhiên (RF), hồi quy logistic (LR), máy học cực đoan (ELM), mạng perceptron nhiều lớp (MLP), v.v. [5]. Khi áp dụng trực tiếp cho dữ liệu GSA gốc, ngoài chi phí tính toán cao, vấn đề quá khớp do đặc trưng có số chiều lớn và số lượng mẫu nhỏ sẽ làm giảm hiệu suất phân loại.

Truyền thống, để giải quyết vấn đề quá khớp, nhiệm vụ phân loại dữ liệu GSA thường được thực hiện sau khi giảm chiều đặc trưng [6]. Trong [7] và [8], các phương pháp trích xuất đặc trưng cho dữ liệu GSA được trình bày, chẳng hạn như trích xuất từ đường cong chuỗi thời gian, tham số khớp đường cong, miền biến đổi, không gian pha, học thích nghi, v.v. Các phương pháp chọn lọc đặc trưng cũng thường được sử dụng để loại bỏ các đặc trưng dư thừa [9], [10], [11]. Một phương pháp giảm chiều phù hợp có thể làm giảm tình trạng quá khớp của bộ phân loại phía sau và cải thiện hiệu quả phân loại. Tuy nhiên, theo hiểu biết của chúng tôi, không có quy tắc rõ ràng nào để chọn các phương pháp này [12].

Đặc biệt, việc trích xuất đặc trưng đại diện cho dữ liệu GSA dưới các loại khí nồng độ thấp là khó khăn hơn do biên độ phản hồi của cảm biến yếu. Để có kết quả nhận dạng tốt nhất, cần phải thử nghiệm nhiều phương pháp khác nhau để tìm ra phương pháp tốt nhất cho ứng dụng cụ thể, điều này thường đòi hỏi rất nhiều thời gian và công sức. Trong khi đó, sau khi giảm chiều đặc trưng, một số đặc trưng có lợi cho phân loại có thể bị mất.

Ngoài ra, học sâu (deep learning) đã được sử dụng rộng rãi trong các bài toán phân loại dữ liệu cảm biến nhờ khả năng học đặc trưng tự động [13]–[19]. Trong [20], một mạng LeNet-5 cải tiến đã được sử dụng để phân loại dữ liệu GSA và đạt độ chính xác tốt hơn so với SVM, MLP và mạng nơ-ron xác suất. Trong [21], một khung học sâu gồm lớp LSTM tích chập và lớp hồi quy được sử dụng để dự đoán các mô tả mùi và đạt hiệu suất tốt hơn so với việc sử dụng mô hình thời gian hoặc không gian đơn lẻ. Tuy nhiên, học sâu thường cần số lượng mẫu lớn để huấn luyện các tham số của mô hình, nên không phát huy hiệu quả đối với dữ liệu có số lượng mẫu nhỏ.

Gần đây, học tổ hợp (ensemble learning) đã thành công trong nhiều trường hợp nhờ khả năng chuyển đổi các mô hình yếu thành mô hình mạnh [22]–[25]. Trong [24], một mô hình Adaboost hai lớp được tối ưu hóa nhằm cải thiện độ chính xác và độ tin cậy của việc nhận diện nhiều loại dược liệu, và mô hình tổ hợp đạt kết quả tốt nhất so với các bộ phân loại đơn lẻ. Trong [25], một mạng nơ-ron nhân tạo (ANN) được huấn luyện độc lập trên từng cảm biến để nhận diện các hợp chất hữu cơ bay hơi (VOC), sau đó được tổ hợp thành một mô hình, cho thấy cải thiện đáng kể về độ chính xác so với bộ phân loại tốt nhất đơn lẻ.

Không giống các phương pháp truyền thống chỉ huấn luyện một mô hình học đơn, các phương pháp học tổ hợp huấn luyện nhiều mô hình cơ sở rồi tổ hợp chúng để cải thiện hiệu quả học tổng thể. Lấy cảm hứng từ ý tưởng này, chúng tôi đề xuất phương pháp học tổ hợp đặc trưng dựa trên ELM (FE-ELM) để giải quyết bài toán phân loại dữ liệu GSA. Lý do sử dụng ELM làm mô hình cơ sở là vì ELM có ưu điểm về lời giải tối ưu toàn cục và chi phí thời gian thấp [26].

Trong FE-ELM, chuỗi thời gian từ từng cảm biến được giảm mẫu trước, sau đó các đặc trưng giảm mẫu từ các cảm biến khác nhau được kết hợp để tạo thành các tập hợp đặc trưng hợp nhất. Tiếp theo, mỗi mô hình ELM cơ sở được huấn luyện riêng trên một tập hợp đặc trưng hợp nhất. Kết quả cuối cùng của FE-ELM cho mẫu kiểm tra được đưa ra bằng cách bỏ phiếu các dự đoán từ tất cả các mô hình cơ sở.

So với các phương pháp truyền thống, phương pháp đề xuất có những ưu điểm sau:

1. So với các bộ phân loại đơn cổ điển, FE-ELM huấn luyện nhiều mô hình cơ sở trên các tập con đặc trưng và kết hợp chúng thành một mô hình mạnh, giúp vượt qua vấn đề quá khớp và có khả năng phân loại mạnh mẽ hơn.
2. So với các phương pháp giảm chiều trước khi phân loại, FE-ELM có thể áp dụng trực tiếp cho dữ liệu GSA mà không làm mất thông tin ban đầu, do đó đơn giản và hiệu quả hơn.
3. So với các phương pháp học tổ hợp truyền thống (huấn luyện mô hình cơ sở trên các tập con mẫu), FE-ELM huấn luyện trên các tập con đặc trưng hợp nhất, khắc phục vấn đề quá khớp của mô hình.

Phương pháp đề xuất được đánh giá trên một tập dữ liệu GSA tự xây dựng với khí nồng độ thấp và một tập dữ liệu công khai. Kết quả thực nghiệm cho thấy phương pháp vượt trội hơn so với các phương pháp truyền thống với hiệu suất phân loại hiệu quả và mạnh mẽ hơn.

**II. THU THẬP DỮ LIỆU VÀ TIỀN XỬ LÝ**

Phần lớn dữ liệu từ mảng cảm biến khí (GSA) được thu thập khi phản hồi với các khí ở nồng độ ppm. Tuy nhiên, kết quả phát hiện của GSA với nồng độ khí thấp hơn vẫn chưa được hiểu rõ. Trong phần này, chúng tôi đã thu thập dữ liệu phản hồi của GSA đối với các loại khí ở mức nồng độ thấp (đơn vị ppb).

**A. Nền tảng thí nghiệm**

Nền tảng thí nghiệm bao gồm hệ thống kiểm tra phần tử cảm biến khí (WS-30B) do hãng Winsen (Trung Quốc) sản xuất và một mảng cảm biến khí. Hình 1(a) là ảnh chụp của nền tảng thí nghiệm. Hệ thống kiểm tra bao gồm một máy tính, mô-đun thu thập dữ liệu, hộp chứa khí và các thành phần khác. Nó cung cấp điện áp sưởi (Vh) và điện áp mạch (Vc) cần thiết cho cảm biến khí hoạt động và hỗ trợ kiểm tra đồng thời tối đa 32 cảm biến.

Khi khí được đưa vào hộp, điện áp Vout trên điện trở tải R1 nối tiếp với cảm biến sẽ được thu thập bởi card thu thập dữ liệu AD và truyền tới phần mềm máy tính để hiển thị dưới dạng đồ thị. Các tín hiệu điện áp này phản ánh đặc điểm phản hồi của cảm biến với khí.

Hình 1(b) minh họa sơ đồ mạch cho một cảm biến duy nhất. Mảng cảm biến bao gồm 10 cảm biến bán dẫn ôxít kim loại khác nhau, có khả năng phản hồi phổ rộng và phát hiện nhiều loại hợp chất hữu cơ dễ bay hơi (VOCs) như: rượu, xeton, benzen, este, ankan, v.v. [27].

Dựa trên các tiêu chí an toàn, ổn định, tính phổ biến và dễ thu thập, chúng tôi đã chọn 6 loại khí tiêu biểu: ethanol, acetone, toluen, ethyl acetate, isopropanol, và n-hexane. Thuật toán được xây dựng dựa trên dữ liệu phản hồi của mảng cảm biến với 6 loại khí này có tính ứng dụng và khả năng mở rộng cao. Thông tin chi tiết về các cảm biến khí được trình bày trong Bảng I.

**B. Chuẩn bị khí tiêu chuẩn**

Trước tiên, chúng tôi chuẩn bị khí tiêu chuẩn ở nồng độ 30 ppm. Sáu túi Teflon dung tích 20 lít được bơm đầy không khí tinh khiết, sau đó lần lượt bơm một lượng phù hợp các mẫu lỏng: ethanol, acetone, toluene, ethyl acetate, isopropanol, và n-hexane bằng bơm tiêm vi lượng. Sau khi chất lỏng bay hơi hoàn toàn, thu được 6 loại khí có nồng độ 30 ppm.

Lượng chất lỏng cần tiêm (Vx, đơn vị µL) được tính theo công thức:



Trong đó:

* ddd là độ tinh khiết của chất lỏng,
* ppp là khối lượng riêng (g/ml),
* VmV\_mVm​ là thể tích mol khí ở nhiệt độ và áp suất chuẩn (24.5 L/mol),
* MMM là khối lượng phân tử tương đối (g/mol),
* C=30C = 30C=30 là nồng độ khí cần tạo (ppm),
* V=20V = 20V=20 là thể tích túi khí (lít).

Bảng II trình bày các thông số kỹ thuật và lượng chất hóa học được sử dụng để cấu hình khí tiêu chuẩn.

Sau đó, chúng tôi tiêm 30, 60 và 120 ml khí tiêu chuẩn vào hộp có thể tích 18 lít để pha loãng thêm. Khi đó, nồng độ khí trong hộp sẽ tương ứng là 50, 100 và 200 ppb.

**C. Thu thập dữ liệu**

Chúng tôi bắt đầu thu thập dữ liệu như sau:

1. Bật nguồn của hệ thống kiểm tra cảm biến khí.
2. Đặt thông số thí nghiệm:
   * Điện áp sưởi và điện áp mạch: 5V;
   * Thời gian đường nền, hấp phụ và khử hấp phụ: lần lượt là 5, 10 và 15 phút;
   * Tần suất lấy mẫu: 1 Hz.

Quy trình thu thập dữ liệu gồm 3 bước:

1. Thu thập dữ liệu phản hồi nền (5 phút) trong không khí nền trong hộp.
2. Tiêm khí tiêu chuẩn vào hộp bằng ống tiêm và thu thập dữ liệu phản hồi (10 phút).
3. Mở hộp để làm sạch cảm biến, giải phóng khí hấp phụ.

Thực hiện quy trình này cho 6 loại khí, mỗi loại ở 3 mức nồng độ khác nhau, lặp lại 5 lần mỗi mức. Kết quả thu được 90 mẫu, mỗi mẫu bao gồm 10 chuỗi thời gian, mỗi chuỗi dài 900 điểm, tương đương 9000 đặc trưng.

**D. Tiền xử lý dữ liệu**

Do tín hiệu phản hồi gốc thường chứa nhiễu tần số cao, chúng tôi áp dụng bộ lọc thông thấp FIR cấp 100, tần số cắt 0.01 để loại bỏ nhiễu.

Ví dụ, với cảm biến TGS2603 phản hồi khí ethanol ở 3 nồng độ khác nhau, Hình 2 trình bày đường cong phản hồi gốc và các đường cong đã giảm mẫu với chu kỳ lấy mẫu 32, 64 và 128.

**III. PHƯƠNG PHÁP ĐỀ XUẤT**

Trong phần này, chúng tôi sẽ giới thiệu ngắn gọn về học tổ hợp (ensemble learning) và máy học cực đoan chuẩn (ELM), sau đó trình bày chi tiết phương pháp ELM tổ hợp đặc trưng (Feature Ensemble-based ELM - FE-ELM) nhằm giải quyết bài toán phân loại dữ liệu GSA có số chiều cao và kích thước mẫu nhỏ. Tiêu chí đánh giá hiệu suất của phương pháp này trên tập dữ liệu nhỏ cũng được trình bày.

**A. Học tổ hợp**

Gần đây, các phương pháp học tổ hợp đã nhận được rất nhiều sự quan tâm, không chỉ vì đạt được kết quả xuất sắc trong các cuộc thi, mà còn được ứng dụng thành công trong hầu hết các lĩnh vực học máy [28]. Thành công của học tổ hợp đến từ khả năng biến các mô hình học yếu thành mô hình mạnh, trong khi mô hình học yếu thì lại dễ tạo ra trong thực tế.

Hai công trình tiên phong [29], [30] không chỉ cho thấy thực nghiệm rằng việc tổ hợp nhiều bộ phân loại thường cho kết quả chính xác hơn so với từng bộ phân loại đơn lẻ, mà còn chứng minh được điều đó về mặt lý thuyết.

Các phương pháp học tổ hợp có thể chia thành hai dạng chính: **boosting** và **bagging**.

* **Boosting**: huấn luyện tuần tự các mô hình cơ sở, trong đó các mẫu bị dự đoán sai ở bước trước sẽ được chú ý nhiều hơn ở bước sau, sau đó kết hợp các mô hình này thành một mô hình mạnh.
* **Bagging**: huấn luyện song song nhiều mô hình cơ sở trên các tập con mẫu khác nhau (lấy mẫu lại bằng phương pháp bootstrap), sau đó gộp kết quả dự đoán bằng trung bình hoặc bỏ phiếu.

**Boosting** giải quyết hiện tượng học thiếu (underfitting) bằng cách giảm sai số huấn luyện dựa trên mối tương quan giữa các mô hình cơ sở. **Bagging** cải thiện khả năng tổng quát hóa bằng cách giảm phương sai dự đoán dựa trên tính độc lập của các mô hình cơ sở.

Giả sử f(x)f(x)f(x) là hàm phân biệt thực và h(x)h(x)h(x) là mô hình cơ sở huấn luyện trên tập con D lấy từ mẫu ban đầu bằng bootstrap. Mô hình tổ hợp từ bagging là:

Theo bất đẳng thức:

→ lỗi bình phương trung bình của tổ hợp nhỏ hơn lỗi của mô hình đơn. Tuy nhiên, các bộ phân loại cơ sở trong cả boosting và bagging đều được huấn luyện trên các tập mẫu khác nhau, dẫn đến vấn đề **quá khớp** trên dữ liệu GSA.

**B. Máy học cực đoan (Extreme Learning Machine - ELM)**

ELM được đề xuất trong [26] để huấn luyện mạng nơ-ron truyền thẳng một lớp ẩn (SLFN). Mạng SLFN gồm một lớp đầu vào, một lớp ẩn và một lớp đầu ra.

Trong ELM:

* Trọng số và bias của lớp ẩn được khởi tạo ngẫu nhiên và **giữ nguyên** trong suốt quá trình huấn luyện.
* Duy nhất cần học là trọng số giữa lớp ẩn và lớp đầu ra.

Mặc dù các nút lớp ẩn được khởi tạo ngẫu nhiên, ELM vẫn duy trì khả năng xấp xỉ hàm toàn cục (universal approximation). So với các phương pháp tối ưu dựa trên gradient truyền thống, ELM có thể tìm được nghiệm tối ưu toàn cục với **chi phí tính toán thấp** nhờ giải bài toán bình phương tối thiểu.

Với những ưu điểm đó, ELM rất phù hợp để làm mô hình cơ sở trong học tổ hợp.

**C. ELM tổ hợp đặc trưng (Feature Ensemble-based ELM - FE-ELM)**

Lấy cảm hứng từ bagging, chúng tôi đề xuất **FE-ELM** để giải quyết bài toán phân loại dữ liệu GSA. Kiến trúc của FE-ELM được trình bày trong Hình 3.

FE-ELM bao gồm ba bước:

1. **Giảm mẫu** chuỗi thời gian của từng cảm biến với cùng chu kỳ và điểm bắt đầu.
2. **Hợp nhất đặc trưng** đã giảm mẫu của nhiều cảm biến để tạo ra một tập đặc trưng hợp nhất.
   * Bằng cách thay đổi điểm bắt đầu giảm mẫu, có thể tạo ra nhiều tập hợp đặc trưng hợp nhất khác nhau.
3. **Huấn luyện mô hình ELM cơ sở** trên từng tập đặc trưng hợp nhất.
   * Kết quả cuối cùng được xác định bằng cách **bỏ phiếu** từ tất cả các mô hình cơ sở.

**Ký hiệu:**

* dữ liệu gốc, gồm mmm cảm biến, mỗi cảm biến có chuỗi thời gian dài ddd, và nnn mẫu.
* : nhãn phân loại cho từng mẫu.
* Sau mã hóa one-hot → (với ccc là số lớp).

**Quy trình huấn luyện:**

1. Thực hiện giảm mẫu với chu kỳ và điểm bắt đầu t (1 ≤ t ≤ τ).
2. Kết hợp đặc trưng của từng cảm biến lại tạo thành tập đặc trưng (với
3. Tạo được τ\tauτ tập dữ liệu huấn luyện tương ứng: X1,X2,...,XτX^1, X^2, ..., X^\tauX1,...,Xτ.
4. Với mỗi t, khởi tạo trọng số và , với h là số nút lớp ẩn.
5. Tính đầu ra lớp ẩn:
6. Tính trọng số lớp đầu ra bằng công thức bình phương tối tiểu:

với (Ht)+(H^t)^+(Ht)+ là nghịch đảo Moore-Penrose.

**Quy trình kiểm tra:**

1. Chia mẫu kiểm tra xxx thành τ\tauτ phần như trong huấn luyện.
2. Với mỗi phần, tính đầu ra:
3. Lấy nhãn dự đoán:
4. Kết quả cuối cùng được bỏ phiếu theo đa số:

**Độ phức tạp thời gian:**

* Mỗi ELM cơ sở có độ phức tạp:

với v=min⁡(n,h)v = \min(n, h)v=min(n,h)

* Tổng độ phức tạp:

**Tiêu chí đánh giá:**

* Độ chính xác:
* Đánh giá bằng **k-fold cross-validation**, lặp lại **r lần** để giảm ngẫu nhiên:
  + Trung bình:
  + Độ lệch chuẩn:
  + Khoảng tin cậy 95%:

**IV. KẾT QUẢ THỰC NGHIỆM**

Trong phần này, chúng tôi so sánh phương pháp FE-ELM được đề xuất với các bộ phân loại kinh điển khác dựa trên cả đặc trưng gốc và đặc trưng đã giảm chiều từ hai tập dữ liệu GSA. Đồng thời, chúng tôi cũng trình bày tác động của các tham số lên hiệu suất của phương pháp đề xuất.

**A. Tập dữ liệu và trực quan hóa**

Chúng tôi đánh giá phương pháp FE-ELM trên hai tập dữ liệu:

1. **GSA-LC**: Tập dữ liệu tự thu thập từ mảng cảm biến khí khi phản ứng với khí nồng độ thấp (ppb).
2. **GSA-FM**: Tập dữ liệu công khai từ UCI, được thu thập từ mảng 16 cảm biến ôxít kim loại dưới điều kiện điều biến dòng khí [3]. Gồm bốn loại khí: không khí, acetone, ethanol và hỗn hợp acetone–ethanol, với tổng cộng 58 mẫu.

* Mỗi mẫu GSA-FM gồm 16 chuỗi thời gian, mỗi chuỗi có 7481 điểm.
* Cả hai tập dữ liệu đều có đặc điểm: **số chiều rất cao và số lượng mẫu nhỏ**.

Hình 4 minh họa kết quả giảm chiều bằng PCA cho hai tập dữ liệu, cho thấy sự phức tạp và chồng lấn giữa các lớp.

**B. Thiết lập thí nghiệm**

* Sử dụng **cross-validation 5 lần 5-fold**: chia tập dữ liệu thành 5 phần, mỗi phần luân phiên làm tập kiểm tra.
* Với mỗi bộ phân loại, thực hiện 20 lần cross-validation và tính **trung bình kết quả**.

**Cài đặt các tham số:**

* **KNN**: k∈{1,3,5,7}k \in \{1, 3, 5, 7\}k∈{1,3,5,7}
* **Adaboost**: số cây con ∈{10,20,30,40,50,60}\in \{10, 20, 30, 40, 50, 60\}∈{10,20,30,40,50,60}
* **RF**: số cây ∈{100,200,400,800,1600,3200}\in \{100, 200, 400, 800, 1600, 3200\}∈{100,200,400,800,1600,3200}
* **SVM**: giá trị log của tham số CCC và tham số kernel G∈[−4,4]G \in [-4, 4]G∈[−4,4]
* **ELM**: số nút lớp ẩn ∈{200,400,600,800,1000}\in \{200, 400, 600, 800, 1000\}∈{200,400,600,800,1000}
* **FE-ELM**: lấy chu kỳ giảm mẫu là 64, số nút lớp ẩn là 200.

Dữ liệu đầu vào cho FE-ELM không được chuẩn hóa vì các đặc trưng có cùng mức độ, nhưng với đặc trưng trích xuất thì áp dụng **chuẩn hóa Z-score**.

Toàn bộ thí nghiệm chạy trên máy tính Intel i7-9700, Windows 10, Python 3.7.

**C. Đánh giá ảnh hưởng của tham số**

Hình 5(a–c) cho thấy ảnh hưởng của **kích thước lớp ẩn** với các chu kỳ giảm mẫu khác nhau. FE-ELM khá ổn định với thay đổi của tham số này trong khoảng hợp lý nhờ tính tổ hợp mô hình.

Hình 5(d–f) minh họa ảnh hưởng của **chu kỳ giảm mẫu**. Khi tăng chu kỳ:

* Ban đầu, độ chính xác tăng vì giảm số chiều giúp tránh quá khớp.
* Nhưng nếu chu kỳ quá lớn → số đặc trưng còn lại quá ít → hiệu suất giảm.

**D. So sánh các bộ phân loại trên dữ liệu gốc**

Chúng tôi so sánh **FE-ELM** với:

* **KNN**
* **Adaboost**
* **Random Forest (RF)**
* **Logistic Regression (LR)**
* **MLP**
* **ELM**
* **SVM tuyến tính (SVM-L)**
* **SVM với kernel (SVM-K)**

**Kết quả:**

* Hình 6: Boxplot cho thấy FE-ELM vượt trội về độ chính xác trung bình, phần lớn hơn tất cả phương pháp khác.
* **Kiểm định t một phía**: so sánh FE-ELM với từng phương pháp khác, kết quả ở Bảng III. Với mức ý nghĩa 0.05:
  + Tất cả p-value đều gần 0 → chứng minh kết quả FE-ELM **có ý nghĩa thống kê cao**.
  + Ngoại lệ: LR và SVM-L trên GSA-FM có p-value lần lượt là 0.1681 và 0.2388.

Bảng IV trình bày **độ chính xác trung bình, khoảng tin cậy 95% và thời gian chạy trung bình** cho từng phương pháp. FE-ELM vừa có độ chính xác cao, vừa tiết kiệm thời gian.

**Ma trận nhầm lẫn:**

Hình 7 là ma trận nhầm lẫn kết quả phân loại của FE-ELM:

* Trên GSA-LC: mỗi lớp có 15 mẫu, tổng cộng 300 mẫu kiểm tra sau 20 lần 5-fold. Tỷ lệ dự đoán đúng từng lớp: 275, 273, 298, 293, 283, 278.
* Trên GSA-FM: có 4 lớp, số mẫu kiểm tra lần lượt là 200, 300, 300, 400. Số mẫu dự đoán đúng là 200, 296, 295, 290.

→ Các ô đường chéo đậm hơn hẳn, chứng tỏ **độ chính xác rất cao**:  
≥91% trên GSA-LC và ≥96.75% trên GSA-FM.

**E. So sánh với SVM trên đặc trưng giảm chiều**

Chúng tôi cũng so sánh **FE-ELM trên đặc trưng gốc** với **SVM-K trên đặc trưng trích xuất** từ 6 phương pháp:

1. **MMB** – độ cao đỉnh phản hồi trừ đường nền [10]
2. **CPC** – hệ số đa thức bậc 3 khớp với đường cong
3. **DFT** – hệ số biến đổi Fourier rời rạc [32]
4. **RDM** – giá trị chênh lệch tương đối cực đại [33]
5. **LDA** – phân tích tuyến tính rời rạc [34]
6. **AEN** – mạng autoencoder [35]

**Kết quả (Bảng V):**

* FE-ELM đạt **độ chính xác cao nhất**, vượt xa SVM-K trên các đặc trưng trích xuất.
* Nguyên nhân: **giảm chiều có thể làm mất thông tin hữu ích**, trong khi FE-ELM sử dụng **gần như toàn bộ đặc trưng gốc**.

→ FE-ELM có thể **áp dụng trực tiếp lên dữ liệu gốc**, thuận tiện hơn nhiều so với các phương pháp truyền thống cần giảm chiều trước rồi mới phân loại.

**V. KẾT LUẬN**

Trong bài báo này, chúng tôi đã đề xuất một khung phương pháp **FE-ELM** (Feature Ensemble-based Extreme Learning Machine) để phân loại dữ liệu từ mảng cảm biến khí (GSA). Khác với các phương pháp truyền thống thực hiện phân loại sau khi giảm chiều đặc trưng, **FE-ELM có thể áp dụng trực tiếp** lên dữ liệu GSA gốc.

Các mô hình cơ sở trong FE-ELM được huấn luyện trên các tập hợp đặc trưng hợp nhất thu được bằng cách **giảm mẫu và kết hợp đặc trưng từ nhiều cảm biến**, giúp:

* **Khắc phục hiện tượng quá khớp** thường gặp trong phân loại dữ liệu GSA.
* **Nâng cao hiệu suất và tính ổn định của mô hình** nhờ chiến lược học tổ hợp.

Kết quả thực nghiệm cho thấy:

* Phương pháp FE-ELM **vượt trội hơn rõ rệt** so với các phương pháp truyền thống về độ chính xác phân loại.
* Hiệu quả được chứng minh không chỉ trên **tập dữ liệu công khai**, mà còn trên **tập dữ liệu tự xây dựng với khí nồng độ thấp** – điều này cho thấy **tiềm năng của FE-ELM trong việc mở rộng giới hạn phát hiện của mảng cảm biến khí**.

**Định hướng tương lai**

Trong tương lai, chúng tôi sẽ tiếp tục phát triển phương pháp đề xuất **trên dữ liệu phản hồi cảm biến đối với các hợp chất hữu cơ dễ bay hơi (VOCs) phức tạp hơn**, nhằm ứng dụng rộng rãi hơn trong thực tế.