DỰ ĐOÁN MỨC ĐỘ BỤI $PM_{2.5}$ BẰNG PHƯƠNG PHÁP KHAI PHÁ DỮ LIỆU

Nguyễn Quỳnh Chi*
* Học Viện Công Nghệ Bưu Chính Viễn Thông

***Tóm tắt*—**Tình trạng ô nhiễm không khí trên toàn cầu không ngừng gia tăng và gây ra những tác động tiêu cực tới sức khỏe con người như: các bệnh đường hô hấp, tim mạch và ung thư. Tại Hà Nội, trong thời gian gần đây, tình hình ô nhiễm càng trở nên xấu hơn, đặc biệt là mật độ bụi $PM_{2.5}$ luôn ở mức cao. Vì vậy, việc dự đoán mức độ ô nhiễm của chỉ số $PM_{2.5}$ trở nên cần thiết hơn nhằm thực hiện cảnh báo sớm. Với dữ liệu về không khí gồm các chỉ số khí tượng và các chỉ ô nhiễm không khí thu thập được tại Hà Nội, chúng tôi thực hiện một phương pháp trích rút đặc trưng mới cho kết quả tốt hơn khi chạy cùng một thuật toán so với phương pháp cũ. Thuật toán XGBoost được áp dụng để dự đoán mức độ ô nhiễm của bụi $PM_{2.5}$ và thử nghiệm đã cho thấy độ chính xác của thuật toán này cao hơn với so với các thuật toán khai phá dữ liệu khác trong khi thời gian huấn luyện lại thấp hơn đáng kể.

***Từ khóa*—** dự đoán chất lượng không khí, khai phá dữ liệu, dự đoán bụi $PM_{2.5}$, XGBoost.

I. GIỚI THIỆU

Tình trạng ô nhiễm không khí gia tăng đang làm phát sinh nhiều vấn đề tới sức khỏe của con người. Theo thông tin được đăng tải bởi Tổ chức Y tế thế giới (WHO), vấn đề ô nhiễm không khí ảnh hưởng tới tất cả mọi người ở các quốc gia [1]. Điều này gây ra 4,2 triệu người chết sớm trên phạm vi toàn cầu trong năm 2016. Trong đó, các nước ở khu vực Đông Nam Á và Tây Thái Bình Dương chiếm 91%. Nguyên nhân chủ yếu đến từ các hạt bụi mịn có kích thước 2,5 µm hoặc nhỏ hơn có trong ô nhiễm không khí, tác nhân gây ra các bệnh tim mạch, hô hấp và ung thư.

Vấn đề ô nhiễm không khí xảy ra nghiêm trọng hơn tại các thành phố lớn do mật độ dân cư cao khiến lượng phát thải khí tăng lên. Bên cạnh đó, việc thi công các công trình xây dựng, đường cũng khiến làm tăng lượng bụi trong không khí tại các thành phố lớn. Thành phố Hà Nội đang phải đối mặt với tình trạng gia tăng ô nhiễm không khí. Trong những ngày tháng 09/2019, Hà Nội được xếp vào một trong những thành phố ô nhiễm không khí cao nhất thế giới. Nguyên nhân chủ yếu tới từ mật độ bụi $PM_{2.5}$ tăng ở mức cao trong không khí. Loại bụi này tác động tiêu cực tới sức khỏe con người, chính vì vậy, dự đoán mức độ ô nhiễm bụi $PM_{2.5}$ càng trở nên cần thiết.

Trong nhiều năm qua, tại các quốc gia phát triển, có nhiều phương pháp dự đoán ô nhiễm bụi $PM_{2.5}$ đã được nghiên cứu. Các thuật toán được áp dụng như hệ lai kết hợp với suy diễn mờ, rừng ngẫu nhiên (Random Forest-RF), máy vectơ hỗ trợ (Support Vector Machine-SVM) và mạng nơ-ron. Những thuật toán này cho kết quả khả quan về độ chính xác dự đoán. Tuy nhiên, những phương pháp này lại thực hiện trên những tập dữ liệu được thu thập tại những thời điểm và địa điểm khác nhau nên khó có thể chọn ra một phương pháp dự đoán từ những nghiên cứu trên phù hợp với dữ liệu về không khí thu thập được tại thành phố Hà Nội.

Vì vậy, chúng tôi đã thực hiện khảo sát các nghiên cứu khác nhau liên quan tới dự đoán mức độ ô nhiễm của chỉ số $PM_{2.5}$ nhằm có cái nhìn tổng quan về các phương pháp dự đoán trong phần 2. Trên cơ sở đó, trong phần 3 chúng tôi thực hiện phân tích dữ liệu thu thập được, đề xuất cách trích rút đặc trưng mới và lựa chọn phương pháp huấn luyện mô hình phù hợp để dự đoán mức độ ô nhiễm chỉ số $PM_{2.5}$ tại thành phố Hà Nội một tiếng sau đó. Các chỉ số về khí tượng là cần thiết cho việc dự đoán, bên cạnh đó các chỉ số ô nhiễm khác (bụi mịn có đường kính cỡ 10 µm – PM_{10}, nồng độ khí CO_2, tổng vật chất hữu cơ lơ lửng – TVOC) và yếu tố thời gian cũng được xem xét ảnh hưởng tới kết quả dự đoán. Với cách trích rút này, chúng tôi thực hiện việc so sánh với phương pháp trích rút cũ và thử nghiệm với các mô hình dự đoán khác nhau: SVM, RF, Perceptron đa lớp (Multi-layer Perceptron-MLP) và XGBoost (Extreme Gradient Boosting) trong phần 4. Cuối cùng, chúng tôi kết luận và thảo luận về hướng phát triển tiếp theo trong tương lai trong phần 5.

II. KHẢO SÁT

Trong phần này, chúng tôi thực hiện khảo sát các nghiên cứu liên quan. Trước hết, một số nghiên cứu đã áp dụng hệ nơ-ron suy diễn mờ thích (Adaptive Neuro Fuzzy Inference System – ANFIS) để dự đoán. Việc sử dụng ANFIS cho thấy có sự cải thiện khi chỉ sử dụng phương pháp suy diễn mờ quy nạp (Fuzzy Inductive Reasoning – FIR), tuy nhiên, sự khác biệt không quá nhiều. Điều này được chỉ ra bởi nghiên cứu dự đoán mật độ bụi $PM_{2.5}$ tại khu vực trung tâm thành phố Mexico [2]. Tuy nhiên, nghiên cứu này không khai thác nhiều các yếu tố khí tượng

Tác giả liên lạc: Nguyễn Quỳnh Chi,
Email: chinq@ptit.edu.vn
Đến tòa soạn: 24/10/2020, chỉnh sửa: 24/11/2020, chấp nhận đăng: 04/12/2020.

vào việc dự đoán mật độ bụi $PM_{2.5}$. Một nghiên cứu khác thực hiện dự đoán bụi PM_{10} tại thành phố Konya [3] cũng sử dụng ANFIS, họ chỉ dùng các yếu tố khí tượng gồm: nhiệt độ, độ ẩm, áp suất và tốc độ gió trong việc dự đoán. Đặc biệt trong việc xử lý dữ liệu, họ có đề xuất phương pháp tỷ lệ dữ liệu phụ thuộc đầu ra (Output-dependent data scaling-ODDS). Điều này đã cho một kết quả hứa hẹn hơn. Tuy nhiên họ không kết hợp thêm giá trị lịch sử của mật độ bụi PM_{10} để dự đoán. Với bài toán dự đoán theo thời gian, việc lựa chọn phương pháp suy diễn mờ là không phù hợp, bởi kết quả dự đoán cho độ chính xác không cao (35% - 62%). Bên cạnh đó, việc không xét đến yếu tố thời gian cũng khiến việc dự đoán trở nên kém chính xác.

Ngoài ra, có những nghiên cứu áp dụng các thuật toán khác như SVM, RF trong việc dự đoán chất lượng không khí. Những nghiên cứu này đều sử dụng các yếu tố khí tượng và giá trị lịch sử các chất ô nhiễm làm đầu vào cho thuật toán của mình. Với các nghiên cứu sử dụng phương pháp SVM [4] [5], kết quả tuy tốt nhưng với mỗi chất ô nhiễm lại chỉ phù hợp với hàm nhân (kernel) nhất định. Theo kết quả thử nghiệm thì với chỉ số SO_2, hàm nhân RBF cho kết quả tốt nhất nhưng với chỉ số NO_2 thì sử dụng hàm tuyến tính lại cho kết quả tốt nhất [4]. Bên cạnh SVM, RF cũng là thuật toán được một số nghiên cứu áp dụng trong việc xây dựng phương pháp dự đoán chất lượng không khí. Một nghiên cứu được thực hiện tại thành phố Thẩm Dương (Trung Quốc) [6] đã xây dựng thuật toán RAQ dựa trên RF để dự đoán chất lượng không khí trong thành phố. Họ xây dựng và thử nghiệm trên tập dữ liệu thu thập từ 10 trạm quan trắc bao gồm nhiều yếu tố: dữ liệu khí tượng, dữ liệu các chỉ số ô nhiễm không khí, dữ liệu về giao thông và địa lý. Phương pháp dự đoán với thuật toán RAQ cho kết quả vượt trội, độ chính xác lên tới 81.5%, trong khi với mạng nơ-ron nhân tạo (Artificial Neural Network – ANN) chỉ đạt 71.8% và cây quyết định (Decision Tree) chỉ đạt 77.4%. Một nghiên cứu khác cũng áp dụng RF trong phương pháp dự đoán của họ được thực hiện với tập dữ liệu thu thập tại thành phố Warsaw để dự đoán trung bình mức độ ô nhiễm của các chất trong ngày tiếp theo [7]. Phương pháp họ thực hiện gồm 2 giai đoạn chính gồm: lựa chọn đặc trưng và áp dụng phương pháp dự đoán. Với giai đoạn lựa chọn đặc trưng, họ thực hiện 2 phương pháp là sử dụng giải thuật di truyền (Genetic Algorithm – GA) và thử khớp từng bước (Stepwise fit-SF) để loại bỏ bớt đặc trưng từ tập đặc trưng ban đầu. Với giai đoạn dự đoán, họ xây dựng 2 mô hình, một mô hình có các đặc trưng qua các mạng nơ-ron và thuật toán học máy khác (MLP, RBF, SVM) rồi tới RF để tổng hợp kết quả dự đoán của các mạng trên, mô hình còn lại có các đặc trưng là đầu vào trực tiếp cho RF. Các chỉ số ô nhiễm được thử nghiệm để dự đoán trong nghiên cứu này gồm: PM_{10}, SO_2, NO_2, O_3. Kết quả họ thực hiện cho thấy việc lựa chọn đặc trưng có ảnh hưởng tới kết quả dự đoán, phương pháp SF thường cho kết quả cao hơn GA lên tới 2.88%. So với phương pháp sử dụng suy diễn mờ, SVM và RF tỏ ra hiệu quả hơn trong việc dự đoán, cho kết quả dự đoán chính xác hơn. Cách trích rút đặc trưng của những nghiên cứu sử dụng SVM và RF cũng đã xét tới nhiều yếu tố về khí tượng, các chỉ số ô nhiễm, thời gian và cả địa lý. Điều này giúp kết quả dự đoán trở nên chính xác hơn và phù hợp với dữ liệu thu thập được.

Ngoài SVM và RF, mạng nơ-ron cũng được áp dụng trong việc dự đoán bụi $PM_{2.5}$. Nghiên cứu trên tập dữ liệu thu thập được tại Hợp Phì (Trung Quốc) đã cho độ chính xác cao khi dự đoán mật độ bụi $PM_{2.5}$ trong ngày tiếp theo khi sử dụng mạng nơ-ron nhân tạo (ANN) [8]. Dữ liệu của họ bao gồm mật độ bụi $PM_{2.5}$ và dữ liệu về khí tượng. Mô hình thiết kế có vector input gồm: mật độ $PM_{2.5}$ và các yếu tố khí tượng (nhiệt độ, tốc độ gió, hướng gió, độ ẩm). Kết quả nghiên cứu cho dự đoán có độ chính xác cao với các độ đo như sau: Trung bình tuyệt đối lỗi (Mean Absolute Error – MAE) [$\mu g/m^3$]: 0.92472; Căn trung bình bình phương lỗi (Root-mean-square Error – RMSE) [$\mu g/m^3$]: 1.2756; Hệ số xác định (Coefficient of Determination – R^2-score): 0.9188; R: 0.9315. Tuy nhiên, với những nghiên cứu sử dụng SVM và RF được đề cập trước đó thì ANN lại tỏ ra kém hiệu quả hơn. Dù vậy, chúng tôi vẫn cân nhắc thử nghiệm với thuật toán này để giải quyết bài toán của chúng tôi.

Trong những năm gần đây, thuật toán Extreme Gradient Boosting (XGBoost) nổi lên trong việc giải quyết bài toán này. Một số nghiên cứu áp dụng thuật toán này đã cho độ chính xác vượt trội hơn so với RF, MLP với thời gian huấn luyện ngắn hơn [9] [10]. Chính những ưu điểm này mà thuật toán XGBoost được áp dụng ngày càng nhiều trong các bài toán dự đoán bên cạnh các thuật toán học sâu.

III. PHƯƠNG PHÁP THỰC HIỆN

Trong phần này, chúng tôi trình bày về phương pháp thực hiện gồm các bước: phân tích tập dữ liệu thu thập được, đề xuất lựa chọn đặc trưng và xây dựng mô hình dự đoán.

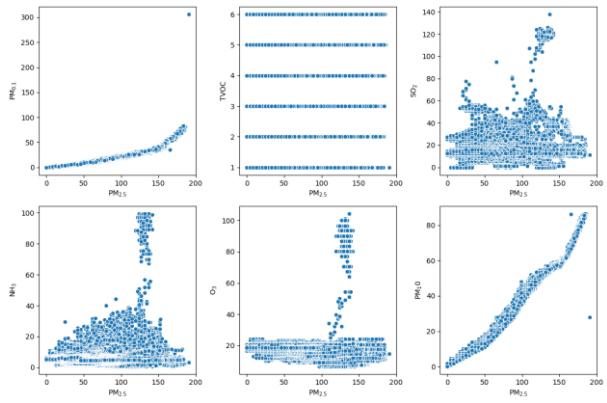
A. Mô tả dữ liệu

Tập dữ liệu của chúng tôi được thu thập tại một trạm quan trắc trong thành phố Hà Nội trong khoảng thời gian từ 17/08/2018 tới 22/07/2019. Mỗi bản ghi trong tập dữ liệu chứa các cột: thời gian, SO_2, NH_3, O_3, $PM_{2.5}$, PM_{10}, CO_2, $PM_{0.1}$, TVOC, CO, nhiệt độ, độ ẩm, ánh sáng. Thời gian lấy mẫu cách nhau trung bình khoảng 40 giây. Tuy nhiên, tập dữ liệu tồn tại một số bản ghi có giá trị rỗng và bị nhiễu. Biểu đồ phân bố các giá trị của thuộc tính (các cột) được mô tả trong *Hình 1*. Sự tồn tại của các bản ghi nhiễu khiến biểu đồ phân bố các giá trị của hầu hết các chỉ số đều bị lệch trái rất nhiều. Tiếp theo chúng tôi thực hiện lọc bỏ các bản ghi nhiễu và trích rút đặc trưng.

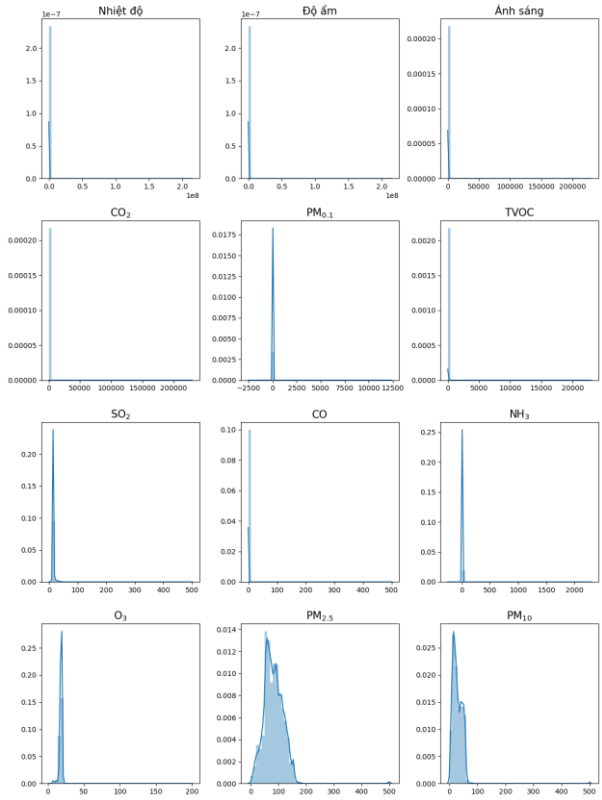
Đầu tiên, chúng tôi thực hiện loại bỏ các bản ghi nhiễu, bị khuyết, mang giá trị nằm ngoài miền cho phép (ví dụ như chỉ số $PM_{0.1}$ tồn tại giá trị âm hoặc nhiệt độ đo được lớn hơn 50 độ C). Qua khảo sát nhiều nghiên cứu [7] [11], các yếu tố về khí tượng: nhiệt độ, độ ẩm, ánh sáng được chúng tôi giữ lại, bởi đây là những chỉ số phản ánh về điều kiện thời tiết và môi trường. Chúng cũng là những nhân tố quan trọng trong mô hình dự đoán mức độ ô nhiễm bụi $PM_{2.5}$.

Tiếp theo, chúng tôi loại bỏ các chỉ số khác không cần thiết bằng cách đánh giá mức độ tương quang với chỉ số $PM_{2.5}$ và giá trị của chị số đó. Dựa trên

và *Bảng II*, có thể thấy rằng chỉ số CO không có ý nghĩa trong việc dự đoán, bởi giá trị của chỉ số này đều bằng 0. Bên cạnh đó những chỉ số TVOC, SO_2, NH_3, O_3 cũng được lược đi bởi chúng không thể hiện được sự tương quan với chỉ số $PM_{2.5}$ như trong



Hình 2.

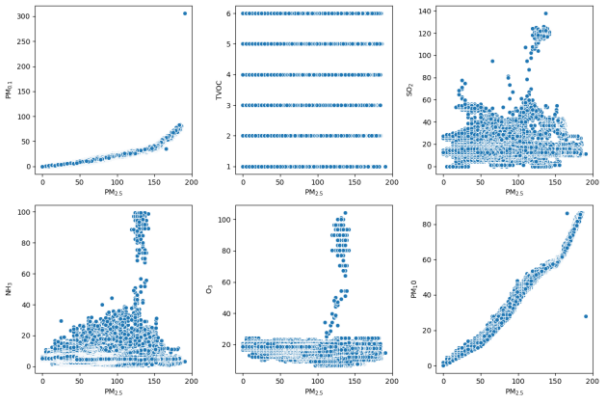


Hình 1 – Phân bố dữ liệu của các chỉ số trong tập dữ liệu

Có thể thấy được rằng chỉ số PM10 và PM2.5 có quan hệ chặt nhất trong số các chỉ số trên nên chỉ số này được giữ lại. Cuối cùng các chỉ số cần thiết để dự đoán: nhiệt độ, độ ẩm, ánh sáng, CO2, PM10 và giá trị phân bố được mô tả như trong Hình 3. Trong phần tiếp theo chúng tôi trình bày phương pháp trích rút đặc trưng từ các chỉ số còn lại sau quá trình tiền xử lý dữ liệu.

Bảng I – Mô tả giá trị các chỉ số CO, NH_3, O_3, PM_{10}

	CO	**NH_3**	**O_3**	**PM_{10}**
Số bản ghi	329455	329455	329455	329455
Trung bình	0.000	5.052	17.991	28.401
Độ lệch chuẩn	0.000	2.374	2.456	15.641
Giá trị nhỏ nhất	0	0.5	6.5	0
25%	0	4.6	17.6	15.7
50%	0	5	18.5	25
75%	0	5.3	18.5	41.7
Giá trị lớn nhất	0	99.6	104.5	86.6



Hình 2 – Biểu đồ tương quan của các chỉ số ô nhiễm với chỉ số $PM_{2.5}$

Bảng II – Mô tả giá trị các chỉ số $PM_{0.1}$, TVOC, SO_2

	$PM_{0.1}$	**TVOC**	**SO_2**
Số bản ghi	329455	329455	329455
Trung bình	17.927	3.535	14.405
Độ lệch chuẩn	8.972	1.605	4.604
Giá trị nhỏ nhất	0	1	0
25%	11	2	12.9
50%	17	4	14.3
75%	24	5	14.3
Giá trị lớn nhất	306	6	137.9

B. *Trích rút đặc trưng*

Các đặc trưng chúng tôi trích ra dựa trên cách lựa chọn đặc trưng của các nghiên cứu chúng tôi đã khảo sát trước đó [7] [8]. Trong đó, phương pháp SF và GA được áp dụng để tìm ra tập các đặc trưng tốt nhất từ tập hợp những đặc trưng ban đầu. Đối với bài toán dự đoán ô nhiễm chỉ số $PM_{2.5}$ tại Hà Nội, chúng tôi thực hiện lấy những đặc trưng tiềm năng được chọn lọc theo kết quả của nghiên cứu tại thành phố Warsaw [7]. Cụ thể hơn, các đặc trưng được đề xuất gồm:

- Các đặc trưng tại thời điểm hiện tại: f_1 – giá trị chỉ số $PM_{2.5}$ hiện tại; f_2 – giá trị chỉ số PM_{10} hiện tại; f_3 – giá trị nhiệt độ hiện tại; f_4 – giá trị độ ẩm hiện tại; f_5 – giá trị ánh sáng hiện tại; f_6 – giá trị chỉ số CO_2 hiện tại. Đây là những giá trị mô tả về không khí ở thời điểm hiện tại nhằm hỗ trợ dự đoán trong giờ tiếp theo.

- Các đặc trưng dựa trên thời gian: f_{7-8} – mùa (biểu diễn bằng 2 bit: 00 – mùa xuân, 01 – mùa hạ, 10 – mùa thu, 11 – mùa đông); f_9 – ngày nghỉ (1 – ngày nghỉ, 0 – ngày đi làm); f_{10} – giờ. Đặc trưng về mùa là cần thiết bởi khí hậu tại Hà Nội là nhiệt đới gió mùa, nên tuy nằm ở khu vực nhiệt đới nhưng lại có 4 mùa thay đổi trong năm. Bên cạnh đó, các đặc trưng về ngày nghỉ trong tuần và thời gian trong ngày cũng được xem xét bởi ô nhiễm không khí chủ yếu do các hoạt động của con người.
- Các đặc trưng trong 24 tiếng trước đó: f_{11-35} – các giá trị chỉ số $PM_{2.5}$ trong 1 tới 24 giờ trước đó. Đặc trưng này phục vụ theo dõi sự biến đổi theo giờ để dự đoán từng giờ tiếp theo.
- Các đặc trưng về khí tượng trong 24 tiếng trước đó: f_{36-38} – giá trị lớn nhất, nhỏ nhất, trung bình chỉ số $PM_{2.5}$ trong 24 giờ trước đó; f_{39-41} – giá trị lớn nhất, nhỏ nhất, trung bình nhiệt độ trong 24 giờ trước đó; f_{42-44} – giá trị lớn nhất, nhỏ nhất, trung bình độ ẩm trong 24 giờ trước đó. Những đặc trưng này nhằm cho thấy mức độ biến động của môi trường trong vòng 24 tiếng, điều này ảnh hưởng tới sự thay đổi của chỉ số $PM_{2.5}$ trong giờ tiếp theo.

So với những nghiên cứu trước đó [7][8], chúng tôi có bổ sung giá trị chỉ số PM_{10} bởi quan sát thấy sự tương quan giữa chỉ số này và giá trị cần dự đoán. Bên cạnh đó, với việc dự đoán theo giờ tiếp theo, việc lấy thêm giá trị lịch sử trong 24 giờ trước đó của chỉ số $PM_{2.5}$ được chúng tôi đưa vào.

Giá trị dự đoán là giá trị trung bình của chỉ số $PM_{2.5}$ trong giờ tiếp theo. Sau khi trích chọn đặc trưng, chúng tôi thực hiện chuẩn hóa dữ liệu bằng chuẩn hóa z – score có công thức (1):

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

Trong đó μ là trung bình các phần tử, σ là độ lệch chuẩn, x là giá trị cần chuẩn hóa.

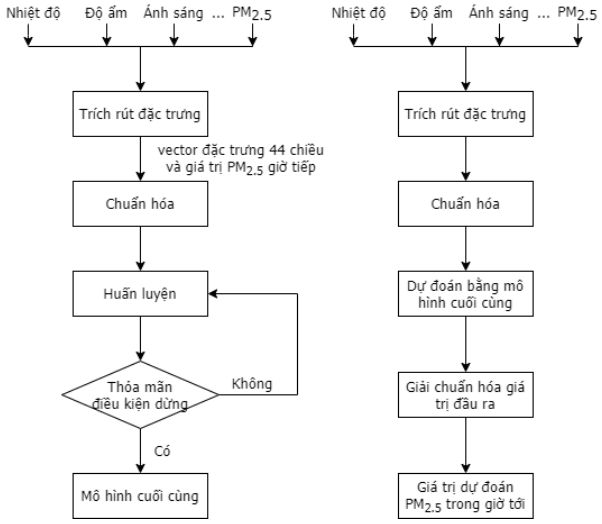
C. Mô tả mô hình dự đoán

Mô hình dự đoán chúng tôi đề xuất gồm quá trình huấn luyện và quá trình dự đoán được trình bày tổng quát trong *Hình 3*.

Với quá trình huấn luyện, từ dữ liệu đầu vào là các chỉ số về khí tượng và các chỉ số ô nhiễm, chúng tôi thực hiện trích rút ra vector đặc trưng 44 chiều như đã trình bày ở phần trước đó. Vector này được chuẩn hóa và thuật toán chúng tôi áp dụng là XGBoost được xây dựng dựa trên Gradient Boost [12]. Khác với RF [13], thuật toán sử dụng phương pháp boosting để giải quyết. Cụ thể hơn, các cây mới được sinh tuần tự với mục đích giảm thiểu lỗi từ cây trước đó bằng việc học lại một phần lỗi từ cây trước đó, cập nhật lỗi để có được cây tốt hơn. Từ đó, tại bước trước, những điểm bị phân sai sẽ có cơ hội được phân đúng nhiều hơn ở xtương lai.

Tập dữ liệu gồm các cặp $(\boldsymbol{x}_i, y_i)$ trong đó $\boldsymbol{x}_i$ là vector đặc trưng 44 chiều và y_i là giá trị dự đoán tương ứng. Mô hình học được mô tả như sau:

$$\hat{y}_i = \phi(\boldsymbol{x}_i) = \sum_{k=1}^{K} f_k(\boldsymbol{x}_i), f_k \in \mathcal{F} \tag{2}$$



Hình 3 – Mô hình dự đoán

Trong đó, $\mathcal{F} = \{f(\boldsymbol{x}) = w_{q(\boldsymbol{x})}\}(q: \mathbb{R}^m) \to T, w \in \mathbb{R}^T$, với q là cây để ánh xạ vector vào giá trị dự đoán tại nút lá, T là số lượng nút lá trên cây, K là số lượng cây, f_k là cây thứ k độc lập trong mô hình, w_i là trọng số của nút lá thứ i và $\hat{y}_i$ là giá trị dự đoán với . Hàm mục tiêu:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k}^{K} \Omega(f_k) \tag{3}$$

Trong đó, n là số điểm dữ liệu, $\Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2$ là hàm qui chuẩn (regularization). Bởi hàm mục tiêu không thể tối ưu bằng phương pháp như Stochastic Gradient Descent (SGD) nên quá trình học được thực hiện như sau: Với $\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(\boldsymbol{x}_i)$ và bắt đầu $\hat{y}_i^{(0)} = 0$, $\hat{y}_i^{(t)}$ là giá trị dự đoán của instance thứ i tại vòng lặp thứ t. Hàm mục tiêu trở thành:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(\boldsymbol{x}_i)) + \Omega(f_t) \tag{4}$$

Và có công thức tính xấp xỉ như sau:

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}^{(t-1)}\right) + g_i f_t(\boldsymbol{x}_i) + \frac{1}{2} h_i f_t^2(\boldsymbol{x}_i) \right] + \Omega(f_t) \tag{5}$$

Với $g_i = \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right)$, $h_i = \partial^2_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right)$. Nếu bỏ phần hằng số, hàm mục tiêu có thể viết đơn giản như sau:

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(\boldsymbol{x}_i) + \frac{1}{2} h_i f_t^2(\boldsymbol{x}_i) \right] + \Omega(f_t) \tag{6}$$

Đặt $G_j = \sum_{i \in I_j} g_i$, $H_j = \sum_{i \in I_j} h_i$, với $I_j = \{i | q(\boldsymbol{x}_i) = j\}$ là tập các giá trị tại nút lá j.

Trọng số tối ưu tại mỗi nút lá:

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{7}$$

Hàm tính lỗi trên toàn bộ cây:

$$\tilde{\mathcal{L}}^{(t)} = -\frac{1}{2}\sum_{j=1}^{T}\frac{{G_j}^2}{H_j + \lambda} + \gamma T \quad (8)$$

Quá trình huấn luyện kết thúc sau một số lần lặp hoặc giá trị hàm mục tiêu nhỏ hơn một ngưỡng nào đó. Mô hình sau khi huấn luyện được sử dụng để dự đoán giá trị trung bình của chỉ số $PM_{2.5}$ trong giờ tiếp theo. Với đầu vào là dữ liệu của các chỉ số khí tượng và ô nhiễm trong vòng 24 tiếng, dữ liệu được trích rút thành một vector 44 chiều sau đó chuẩn hóa. Vector này được đưa vào mô hình đã huấn luyện để đưa ra giá trị dự đoán. Trong phần tiếp theo, chúng tôi thực hiện thử nghiệm phương pháp trích rút và mô hình dự đoán đã được trình bày.

IV. THỬ NGHIỆM

Bởi dữ liệu được chúng tôi thu thập được lấy mẫu cách nhau khoảng 40 giây, nên để thực hiện thử nghiệm, chúng đã lấy trung bình các bản ghi đó theo giờ. Kết quả thu được 6433 bản ghi về các chỉ số không khí theo giờ. Tiếp theo, chúng tôi thực hiện tiền xử lý, trích rút và chuẩn hóa dữ liệu này. Để thực hiện quá trình huấn luyện và đánh giá, các bản ghi được lấy ngẫu nhiên và chia thành 2 tập: tập huấn luyện (training set) chiếm 75% dữ liệu ban đầu và 25% dữ liệu còn lại là tập kiểm tra (test set).

Các độ đo được chúng tôi sử dụng để đánh giá gồm R^2 – score công thức (9), MAE công thức (10) và RMSE công thức (11) như sau:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (9)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (10)$$

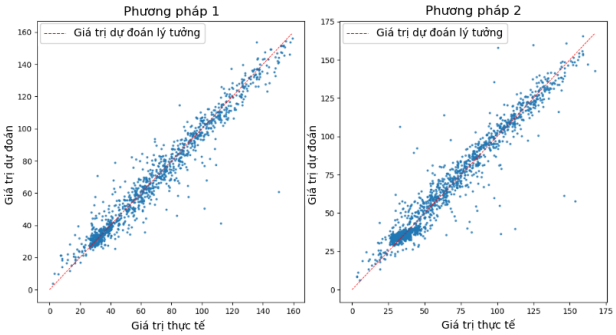
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (11)$$

Trong đó, n là số phần tử, y_i là giá trị thực tế, $\hat{y}_i$ là giá trị dự đoán, $\bar{y}$ là giá trị trung bình của số phần tử. Các độ đo này được sử dụng bởi chúng thể hiện rõ được mức độ chênh lệch giữa giá trị thực tế và giá trị dự đoán. Điều này phù hợp với các bài toán hồi quy (regression) bởi giá trị dự đoán nằm trên miền liên tục thay vì là các nhãn như bài toán phân loại. Đối với R^2 – score, giá trị càng cao thì mô hình càng mạnh (thể hiện mức độ phù hợp với tập dữ liệu) và tốt nhất là 1.00, với MAE và RMSE giá trị càng nhỏ càng tốt (2 độ đo này thể hiện sự sai khác giữa giá trị dự đoán và giá trị thực tế)

Tiếp theo, chúng tôi thực hiện so sánh kết quả giữa phương pháp trích rút của chúng tôi đã trình bày trong phần 3 (Phương pháp 1) và phương pháp trích rút khác chỉ gồm các đặc trưng trích từ các yếu tố khí tượng (Phương pháp 2) [8]. Cụ thể, phương pháp của chúng tôi có xét đến các yếu tố về thời gian trong ngày và trong năm, kèm theo đó là chỉ số PM_{10} và những số liệu đầu vào của các chỉ số trong 24 giờ trước đó, còn với phương pháp 2, họ chỉ quan tâm tới những yếu tố khí tượng trong phạm vi hiện tại. So sánh kết quả thực hiện với các độ đo được trình bày trong *Bảng III* và kết quả dự đoán của 2 phương pháp trong *Hình 4* với bên trái là so sánh giá trị thực tế với giá trị dự đoán khi áp dụng phương pháp 1, bên phải là so sánh giá trị thực tế với giá trị dự đoán khi áp dụng phương pháp 2.

Bảng III – Kết quả so sánh giữa 2 phương pháp

	R^2 – score	MAE	RMSE
Phương pháp 1	**0.9508**	**0.1387**	**0.2266**
Phương pháp 2	0.9368	0.1515	0.2521



Hình 4 – Kết quả dự đoán của 2 phương pháp

Có thể thấy rằng, phương pháp trích rút của chúng tôi cho kết quả cao hơn ~2% so với phương pháp cũ khi thử nghiệm trên cùng một mô hình. Điều này khẳng định rằng yếu tố thời gian và chỉ số PM_{10} có tác động tới kết quả dự đoán chỉ số $PM_{2.5}$ trong giờ tiếp theo bên cạnh những yếu tố cơ bản về khí tượng như: nhiệt độ, độ ẩm, ánh sáng.

Tiếp theo chúng tôi thực hiện so sánh mô hình dự đoán với các mô hình khác: SVM, Random Forest, MLP và XGBoost. Siêu tham số (Hyper-parameter) của mỗi thuật toán được đặt như trong *Bảng IV*.

Bảng IV – Siêu tham số cho mỗi thuật toán

	Hyper-parameter
SVM	gamma='auto' kernel='rbf' C=100 epsilon=0.0001
Random Forest	n_estimators=150 max_features='auto'
MLP	hidden_layer_sizes=(192,128,96) max_iter=1000 learning_rate_init=0.01 tol=1e-6 batch_size=192
XGBoost	n_estimators=200 max_depth=8 gamma=0.7 objective='reg:squarederror'

Các tiêu chí để so sánh tượng tự, gồm các độ đo: R^2 – score, MAE, RMSE và thời gian huấn luyện được tính bằng giây. Kết quả được trình bày trong *Bảng V*.

Bảng V – So sánh kết quả giữa các thuật toán

	R^2 – score	MAE	RMSE	Thời gian
SVM	0.9553	0.1154	0.2101	27.0608

Random Forest	0.9587	0.1115	0.2020	35.5577
MLP	0.9562	0.1276	0.2078	8.2011
XGBoost	**0.9595**	**0.1126**	**0.1999**	**4.8872**

Thông qua độ đo R^2 – score, có thể thấy rằng thuật toán XGBoost cho tỷ lệ phù hợp với tập dữ liệu cao nhất (95,95%). Với kết quả của RMSE, sự chênh lệch giữa giá trị dự đoán và giá trị thực tế là nhỏ nhất, tức độ chính xác của dự đoán là cao nhất khi so với các giá trị còn lại. So sánh với những thuật toán còn lại, tư tưởng của XGBoost là xây dựng các mô hình dự đoán yếu và kết hợp chúng để cho ra mô hình dự đoán cuối cùng có độ chính xác cao. Kết hợp với việc cập nhật lại trọng số bằng phương pháp hạ đạo hàm (gradient descent), thuật toán XGBoost sẽ cho ra mô hình dự đoán khớp với tập dữ liệu nhiều nhất có thể. Tuy sự khác biệt về độ chính xác giữa các thuật toán không quá nhiều nhưng so sánh về thời gian huấn luyện thì XGBoost có thời gian huấn luyện ngắn nhất. Điều này cho thấy tiềm năng của mô hình này trong việc huấn luyện và độ chính xác dự đoán theo thời gian.

V. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Với dữ liệu chúng tôi thu thập được tại Hà Nội gồm các yếu tố về khí tượng và các chỉ số ô nhiễm, chúng tôi đã nhận thấy chỉ số PM_{10} tại Hà Nội có sự tương quan với chỉ số $PM_{2.5}$. Từ đó, cùng với những khảo sát khác chúng tôi thực hiện phương pháp trích rút đặc trưng mới. Phương pháp trích rút mới bao gồm không chỉ các yếu tố về khí tượng và ô nhiễm ở thời điểm hiện tại mà còn trong quá khứ (nhiều giờ trước đó). Điều này giúp dự đoán tốt hơn do giá trị lịch sử giúp thể hiện xu hướng biến đổi của chỉ số $PM_{2.5}$ trong giờ tiếp theo. Ngoài ra, yếu tố về thời gian cũng đóng vai trò tác động lên kết quả dự đoán do sự biến đổi về khí hậu, môi trường theo mùa trong năm tại Hà Nội và hoạt động khác nhau của con người trong từng khung thời gian khác nhau trong ngày và trong tuần. Thử nghiệm đã chứng minh phương pháp trích rút của chúng tôi cho kết quả dự đoán mức độ bụi $PM_{2.5}$ tại Hà Nội tốt hơn so với phương pháp cũ (chỉ quan tâm tới các yếu tố khí tượng).

Nghiên cứu cũng cho thấy thuật toán XGBoost là một thuật toán tốt cho độ chính xác cao với thời gian huấn luyện thấp khi so sánh với các thuật toán học máy khác. Đối với bài toán của chúng tôi, thuật toán này là phù hợp bởi khả năng dự đoán chính xác và chi phí huấn luyện mô hình thấp. Tuy nhiên, bởi tính chất cố gắng khớp với dữ liệu tốt nhất của thuật toán này khiến thuật toán này dễ bị quá mức phù hợp (overfitting). Vì vậy, trong tương lai chúng tôi sẽ xem xét đến một số phương pháp để hạn chế việc bị overfitting và thử nghiệm với các thuật toán học sâu (deep learning) khác để dự đoán cho các bài toán dữ liệu chuỗi thời gian (time-series).

Về mặt dữ liệu hiện tại của chúng tôi cũng thiếu một số yếu tố về khí tượng như hướng gió, tốc độ gió. Đây cũng là những yếu tố có thể ảnh hưởng tới việc dự đoán ô nhiễm không khí do gió có thể khuếch tán hoặc làm tập trung mật độ bụi tại một khu vực nào đó. Với khí hậu tại Hà Nội, gió còn có những đặc trưng khác nhau thay đổi theo mùa như: hướng gió, tốc độ, độ ẩm. Ngoài ra, dữ liệu về giao thông cũng cần được quan tâm bởi lượng phương tiện cá nhân tại Hà Nội rất nhiều. Trong tương lai, chúng tôi sẽ thu thập thêm những dữ liệu trên để quan sát sự tương quan giữa chúng với mức độ ô nhiễm không khí tại Hà Nội và cải tiến hoặc thử nghiệm với mô hình khác nhằm cải thiện độ chính xác, phạm vi dự đoán theo không gian và theo thời gian.

TÀI LIỆU THAM KHẢO

[1] WHO, "Air pollution," 2 May 2018. [Online]. Available: https://www.who.int/en/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health.

[2] À. Nebot and F. Mugica, "Small-particle pollution modeling using fuzzy approaches," *Advances in Intelligent Systems and Computing*, pp. 239-252, 2014.

[3] K. Polat and S. S. Durduran, "Usage of output-dependent data scaling in modeling and prediction of air pollution daily concentration values (PM10) in the city of Konya," *Neural Computing and Applications*, p. 21, 2011.

[4] C.-M. Vong, W.-F. Ip, P.-k. Wong and J.-y. Yang, "Short-Term Prediction of Air Pollution in Macau Using Support Vector Machines," *Journal of Control Science and Engineering*, vol. 2012, 2012.

[5] W.-F. Ip, C.-M. Vong, J. Y. Yang and P. K. Wong, "Least squares support vector prediction for daily atmospheric pollutant level," *Proc. 2010 IEEE/ACIS 9th International Conference on Computer and Information Science (ICIS)*, pp. 23-28, August 2010.

[6] R. Yu, Y. Yang, L. Yang and G. Han, "RAQ–A Random Forest Approach for Predicting Air Quality in Urban Sensing Systems," *Sensors*, vol. 16, p. 86, 11 January 2016.

[7] K. Siwek and S. Osowski, "DATA MINING METHODS FOR PREDICTION OF AIR POLLUTION," *Int. J. Appl. Math. Comput. Sci*, vol. 26, 2016.

[8] A. Li và X. Xu, "A New PM2.5 Air Pollution Forecasting Model Based on Data Mining and BP Neural Network Model," *Advances in Computer Science Rese*, tập 65, 2018.

[9] NandigalaVenkatAnurag, YagnavalkBurra and S.Sharanya, "Air Quality Index Prediction with Meteorological Data Using Feature Based Weighted Xgboost," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, no. 1, pp. 1355-1358, May 2019.

[10] M. Z. Joharestani, C. Cao, X. Ni, B. Bashir and S. Talebiesfandarani, "PM2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data," *Atmosphere*, 2019.

[11] X. Yi, J. Zhang, Z. Wang, T. Li and Y. Zheng, "Deep Distributed Fusion Network for Air Quality Prediction," in *The 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, London, 2018.

[12] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," 2016.

[13] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5-32, 2001.

$PM_{2.5}$ CONCENTRATION PREDICTION BY DATA MINING METHOD

Abstract: The global air pollution is constantly increasing and causing negative effects on human health such as respiratory, cardiovascular diseases and cancers. Recently, pollution in Hanoi has become increasingly worse,

especially when $PM_{2.5}$ concentration is always at high level. Thus, $PM_{2.5}$ prediction is of more urgency to issue early forecasts. Depending on air data including meteorological indicators and air pollution indicators collected in Hanoi, we have proposed a new characteristic extraction method that gave better results when uing the same algorithm compared to those of old methods. XGBoost algorithm was applied to predict the concentration of $PM_{2.5}$ and the test showed that the accuracy of this algorithm is higher than that of other data mining algorithms while the training time is significantly lower.

Keyword: air quality forecasting, data mining, $PM_{2.5}$ prediction, XGBoost

Nguyễn Quỳnh Chi tốt nghiệp đại học chuyên ngành Công nghệ thông tin loại giỏi tại đại học Bách Khoa, Hà nội, Việt nam năm 1999, nhận bằng Thạc Sỹ chuyên ngành Khoa học máy tính tại Đại học California, Hoa Kỳ năm 2004 và nghiên cứu sinh Tiến sỹ Khoa học máy tính từ năm 2004 đến 2008, cũng tại Đại học California, Hoa Kỳ. Lĩnh vực nghiên cứu liên quan tới kho dữ liệu và ứng dụng các phương pháp học máy và khai phá dữ liệu để giải quyết các bài toán trong thực tế