XÂY DỰNG MÔ HÌNH HỌC SÂU ĐÁNH GIÁ NGUY CƠ CHÁY RỪNG TẠI LÂM ĐỒNG

Lê Văn Hưng[1], Nguyễn Thị Thanh[2], Đặng Hữu Nghị[2], Hoàng Anh Đức[2]

[1] *Khoa Công nghệ Thông tin, Trường Đại học Mỏ - Địa chất*, levanhung@humg.edu.vn
[2] *Khoa Công nghệ Thông tin, Trường Đại học Mỏ - Địa chất*,
{nguyenthithanh@humg.edu.vn, danghuunghi@humg.edu.vn, hoanganhduc@humg.edu.vn}

TÓM TẮT

Cháy rừng có thể gây ra những tổn hại nghiêm trọng đến tài sản và đời sống con người, thực vật, động vật, hệ sinh thái và môi trường. Do biến đổi khí hậu cũng như các hoạt động của con người, cháy rừng đã tăng đến mức báo động ở Việt Nam. Dự báo nguy cơ cháy rừng là một yếu tố quan trọng trong phòng cháy, chữa cháy rừng. Trong bài báo này, chúng tôi phát triển một mô hình mạng nơ ron truyền thẳng sâu cho bài toán đánh giá, phân vùng nguy cơ cháy rừng của tỉnh Lâm Đồng. Các yếu tố đầu vào bao gồm độ dốc, hướng dốc, độ cao, hiện trạng sử dụng đất, chỉ số thực vật NDVI, khoảng cách tới đường giao thông, khoảng cách tới khu dân cư, nhiệt độ, tốc độ gió và lượng mưa. Thực nghiệm cho thấy mô hình này cho kết quả rất tốt trên nhiều tiêu chí đánh giá và so với các kỹ thuật học máy khác. Mô hình sau khi xây dựng được sử dụng để tính chỉ số và phân vùng nguy cơ cháy rừng cho vùng nghiên cứu.

Từ khóa: mạng nơ ron truyền thẳng, học sâu, cháy rừng, Lâm Đồng

1. GIỚI THIỆU

Các phương pháp dự báo nguy cơ cháy rừng ở nước ta hiện nay chủ yếu dựa trên các mô hình truyền thống, ví dụ như dựa trên chỉ số tổng hợp P của Nesterov hoặc có cải tiến để phù hợp với điều kiện Việt Nam [3]. Có thể thấy rằng, các phương pháp này đã bỏ qua nhiều yếu tố đầu vào quan trọng cho bài toán dự báo nguy cơ cháy rừng như các chỉ số thảm thực vật, khoảng cách tới đường giao thông, khoảng cách tới khu dân cư ... là những yếu tố đã được chứng minh có khả năng dự báo nguy cơ cháy rừng cao [1]. Nguyễn Ngọc Thạch và nnk [4] đã thành lập bản đồ nguy cơ cháy rừng cho tỉnh Sơn La bằng cách sử dụng phương pháp phân tích thứ bậc (AHP) để xác định trọng số của các yếu tố đầu vào.

Trên thế giới, các phương pháp thống kê đã được sử dụng cho nghiên cứu cháy rừng do tính chất ngẫu nhiên cố hữu của hiện tượng cháy rừng [5]. Tuy nhiên, với các bài toán có khối lượng dữ liệu lớn, nhiều đầu vào, độ chính xác dự báo của các mô hình thống kê vẫn còn hạn chế [1]. Gần đây, các mô hình học máy đã được đề xuất cho cháy rừng do chúng làm việc tốt hơn với dữ liệu lớn, có nhiều đầu vào. Nhìn chung, độ chính xác của các mô hình học máy là tốt hơn các mô hình thống kê [1]. Học sâu, một nhánh nghiên cứu của học máy dựa trên mạng nơ ron nhân tạo, đang trở thành một trong những công cụ cốt lõi của Cuộc cách mạng công nghiệp 4.0. Ở nước ta hiện nay, việc phát triển các mô hình dự báo nguy cơ cháy rừng sử dụng các kỹ thuật học máy hiện đại, đặc biệt là học sâu, còn chưa nhiều. Vì vậy, trong nghiên cứu này, chúng tôi sẽ xây dựng một mô hình mạng nơ ron sâu cho dự báo nguy cơ cháy rừng tại Lâm Đồng, góp phần chứng minh tính hiệu quả của việc áp dụng các mô hình học máy nói chung và học sâu nói riêng cho bài toán đánh giá nguy cơ cháy rừng tại Việt Nam.

2. PHƯƠNG PHÁP

2.1. Thu thập dữ liệu

Trong [1], các tác giả đã xây dựng mô hình học máy PSO-NF đánh giá nguy cơ cháy rừng cho tỉnh Lâm Đồng và công bố kết quả trên một tạp chí quốc tế có uy tín. Do mục tiêu của nghiên cứu này là chứng minh tính hiệu quả của việc sử dụng mô hình học sâu trong đánh giá nguy cơ cháy rừng tại Việt Nam và để tiện cho việc so sánh với các mô hình khác, chúng tôi đã

sử dụng bộ dữ liệu của nghiên cứu nói trên trong bài báo này. Dữ liệu ở định dạng raster của ArcGIS. Đầu vào của mô hình bao gồm 10 yếu tố: Độ dốc, hướng dốc, độ cao, hiện trạng sử dụng đất, chỉ số thực vật NDVI, khoảng cách tới đường giao thông, khoảng cách tới khu dân cư, nhiệt độ, tốc độ gió và lượng mưa. Bộ dữ liệu huấn luyện và kiểm tra mô hình bao gồm 1080 mẫu (điểm/ô), trong đó 756 mẫu được sử dụng để huấn luyện, 324 mẫu còn lại dùng để kiểm tra, đánh giá mô hình và số các điểm cháy rừng và không cháy là bằng nhau.

2.2. Thiết kế kiến trúc mô hình

Mô hình gồm nhiều nơ ron nhân tạo được tổ chức thành các tầng: 01 tầng vào, 01 tầng ra và một hoặc nhiều tầng ẩn [2]. Do bài toán ở đây là học có giám sát với dữ liệu đầu vào là một vec tơ kích thước cố định nên mạng nơ ron truyền thẳng (đầu ra của một tầng là đầu vào của tầng kế tiếp) nhiều tầng ẩn là một mô hình thích hợp. Dữ liệu huấn luyện/kiểm tra mô hình có thể được xem như là một phân phối xác suất Bernoulli của dữ liệu đầu vào, chỉ nhận hai giá trị 0 (lớp không cháy/âm) hoặc 1 (lớp cháy/dương). Mô hình được thiết kế với những lựa chọn sau: (i) Tầng đầu ra: Do ta xem dữ liệu như một phân phối xác suất Bernoulli, 01 nơ ron ở tầng đầu ra với hàm kích hoạt sigmoid được sử dụng, giá trị đầu ra dự báo xác suất mẫu dữ liệu rơi vào lớp cháy. Đầu ra sigmoid thường được sử dụng kết hợp với hàm mất mát (sai số) binary cross-entropy. Hàm mất mát này đo khoảng cách giữa phân phối xác suất của dữ liệu huấn luyện/kiểm tra và của mô hình. Khoảng cách giữa hai phân phối nhỏ đồng nghĩa với việc hai phân phối đó rất gần nhau và mô hình khớp tốt với dữ liệu; (ii) Các tầng ẩn: Hàm kích hoạt sigmoid và tanh thường được sử dụng nhiều cho các nơ ron ẩn trong quá khứ vì có đạo hàm rất đẹp. Những năm gần đây, hàm ReLU được sử dụng rộng rãi vì tính đơn giản, giúp cho việc huấn luyện các mạng nơ ron sâu nhanh hơn rất nhiều. Khi số tầng ẩn và số nơ ron của mỗi tầng ẩn tăng lên, khả năng biểu diễn (khớp với) dữ liệu huấn luyện của mô hình tăng lên. Tuy nhiên, điều này có thể làm mô hình trở nên quá khớp (overfitting), nghĩa là hàm mất mát đạt giá trị rất nhỏ trên tập dữ liệu huấn luyện nhưng lại cao trên tập kiểm tra (khả năng tổng quát hóa kém). Ngược lại, mô hình có thể khớp kém (underfitting) với dữ liệu, nghĩa là có sai số cao trên cả tập huấn luyện và tập kiểm tra. Quá trình thiết kế thường là quá trình thử nghiệm và theo dõi sai số. Ở bài toán này, do số đầu vào và số lượng mẫu huấn luyện không quá lớn, chúng tôi đã thử nghiệm và chọn thiết kế mô hình có 3 tầng ẩn với số nơ ron tương ứng là 20, 10 và 5.

2.3. Chọn thuật toán huấn luyện mô hình

Quá trình huấn luyện điều chỉnh trọng số kết nối giữa các nơ ron để mô hình khớp với dữ liệu huấn luyện. Phương pháp cập nhật trọng số phổ biến nhất là Gradient Descent (GD). Để áp dụng GD, ta cần tính được gradient của hàm mất mát theo từng ma trận trọng số. Phương pháp thường được dùng để tính gradient là lan truyền ngược, tính gradient ngược từ tầng cuối đến tầng đầu tiên. Đối với mạng nơ ron truyền thẳng, các thuật toán cập nhật tối ưu thường được dùng là SGD, RMSprop và Adam. Việc cập nhật trọng số của mô hình có thể được thực hiện theo từng mẫu dữ liệu, cho toàn bộ dữ liệu (batch) hoặc cho một phần dữ liệu (mini-batch). Trong nghiên cứu này, chúng tôi chọn mini-batch GD do thường được sử dụng nhiều nhất.

2.4 Phương pháp đánh giá mô hình

Để đánh giá hiệu năng của một mô hình học máy phân lớp, độ chính xác (tỉ lệ giữa số điểm dự đoán đúng và tổng số điểm) và ma trận confusion matrix thường được sử dụng (Bảng 1). Các tỉ số FNR và FPR còn được gọi tương ứng là *tỉ lệ bỏ sót* và *tỉ lệ báo động nhầm*. Đôi khi, ta có thể chấp nhận tỉ lệ báo động nhầm cao để đạt được tỉ lệ bỏ sót thấp. Việc tăng hay giảm các tỉ lệ này có thể được thực hiện bằng cách thay đổi ngưỡng phân chia giữa lớp dương và lớp âm (mặc định là 0.5). Khi thay đổi ngưỡng từ 0 đến 1, với mỗi một giá trị ngưỡng ta thu được một cặp giá trị (FPR, TPR). Biểu diễn các điểm (FPR, TPR) trên đồ thị ta được đường cong ROC. AUC là diện tích nằm dưới đường cong ROC, cho biết khả năng phân biệt 2 lớp của mô hình và càng lớn càng tốt. Thông thường, mô hình có AUC từ 0.8 trở lên được coi là tốt.

Bảng 1. Ma trận confusion matrix (trái) và ma trận normalized confusion matrix (phải)

	Predicted as Positive	Pred. as Neg.
Actual: Positive	True Positive (TP)	False Negative (FN)
Actual: Negative	False Positive (FP)	True Negative (TN)

	Pred. as Pos.	Pred. as Neg.
Actual: Positive	TPR = TP/(TP + FN)	FNR = FN/(TP + FN)
Actual: Negative	FPR = FP/(FP + TN)	TNR = TN/(FP + TN)

2.5. Huấn luyện mô hình

Để giảm sự quá khớp của mô hình trong quá trình huấn luyện, chúng tôi đã sử dụng kĩ thuật Dropout cho các lớp ẩn (loại bỏ ngẫu nhiên một vài nơ ron). Tỉ lệ Dropout được chọn là 20%.

2.6. Thành lập bản đồ phân vùng nguy cơ cháy rừng cho vùng nghiên cứu

Sau khi mô hình được xây dựng thành công, nó được sử dụng để tính chỉ số nguy cơ cháy rừng cho tất cả các điểm của vùng nghiên cứu. Giá trị đầu ra của mô hình cho mỗi điểm là xác suất điểm đó rơi vào lớp cháy rừng. Các giá trị này được phân thành 6 lớp trên bản đồ, thể hiện các mức độ nguy cơ khác nhau (rất thấp, thấp, bình thường, cao, rất cao và đặc biệt cao).

3. KẾT QUẢ VÀ THẢO LUẬN

Các thử nghiệm cho thấy mô hình được huấn luyện bằng thuật toán cập nhật trọng số Adam, số lượng vòng lặp epoch = 2000 và batch_size = 20 cho kết quả tốt nhất. Kết quả đánh giá hiệu năng của mô hình được thể hiện trong các Bảng 2-4 và Hình 1 sau.

Bảng 2. Các ma trận confusion matrix của tập dữ liệu huấn luyện

	Pred. as Pos.	Pred. as Neg.
Act. Pos.	365	13
Act. Neg.	70	308

	Pred. as Pos.	Pred. as Neg.
Act. Pos.	96.6%	3.4%
Act. Neg.	18.6%	81.4%

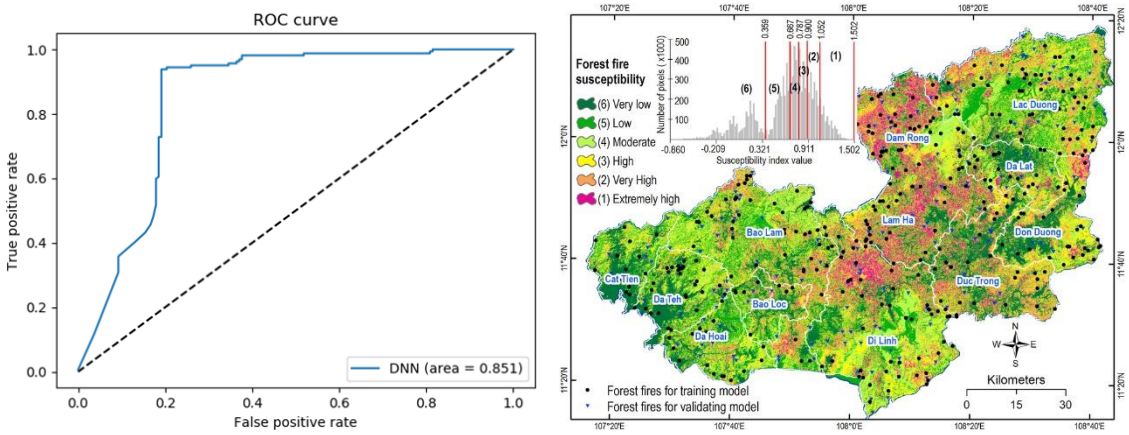
Bảng 3. Các ma trận confusion matrix của tập dữ liệu kiểm tra

	Pred. as Pos.	Pred. as Neg.
Act. Pos.	152	10
Act. Neg.	31	131

	Pred. as Pos.	Pred. as Neg.
Act. Pos.	93.8%	6.2%
Act. Neg.	19.1%	80.9%

Bảng 4. Độ chính xác của mô hình học sâu (DNN) so với các mô hình học máy khác

	DNN	**SVM**	**Random Forests**	**PSO-NF**
Trên tập huấn luyện	89.0%	86.2%	86.4%	89.3%
Trên tập kiểm tra	87.3%	84.9%	85.2%	85.8%



Hình 1. Đường cong ROC và AUC của mô hình DNN trên tập dữ liệu kiểm tra (trái)
Bản đồ phân vùng nguy cơ cháy rừng của tỉnh Lâm Đồng (phải)

Có thể thấy rằng độ chính xác trên tập dữ liệu kiểm tra của mô hình DNN vượt trội so với các mô hình đã thử nghiệm khác, nghĩa là nó có khả năng tổng quát hóa tốt nhất. Trên tập dữ liệu huấn luyện, độ chính xác của mô hình DNN tương đương với mô hình PSO-NF và tốt hơn so với các mô hình SVM và RF. Đối với cả tập dữ liệu huấn luyện và tập dữ liệu kiểm tra, giá trị AUC là tốt (tương ứng là 0.87 và 0.85), đồng thời tỉ lệ bỏ sót tương đối nhỏ (3.4% và 6.2%).

4. KẾT LUẬN

Kết quả nghiên cứu cho thấy mô hình mạng nơ ron truyền thẳng sâu đã xây dựng có khả năng tổng quát hóa tốt nhất trong các mô hình học máy đã thử nghiệm cho bài toán đánh giá, phân vùng nguy cơ cháy rừng của tỉnh Lâm Đồng và do đó hoàn toàn có thể áp dụng cho các vùng nghiên cứu khác tại Việt Nam. Trong thời gian tới, chúng tôi sẽ tiếp tục thử nghiệm mô hình học sâu cho các dạng thiên tai khác như trượt lở đất tại Việt Nam.

TÀI LIỆU THAM KHẢO

[1] Bui, D.T., et al, 2017. A Hybrid Artificial Intelligence Approach Using GIS-Based Neural-Fuzzy Inference System and Particle Swarm Optimization for Forest Fire Susceptibility Modeling at A Tropical Area. *Agricultural and Forest Meteorology*, 233(15), 32–44.

[2] Goodfellow, I., Bengio, Y. & Courville, A., 2016. *Deep Learning*. MIT Press

[3] Phạm Ngọc Hưng, 2004. *Quản lý cháy rừng ở Việt Nam*. Nhà xuất bản Nghệ An.

[4] Nguyễn Ngọc Thạch, Đặng Ngô Bảo Toàn, Phạm Xuân Cảnh, 2017. Ứng dụng viễn thám và GIS thành lập bản đồ nguy cơ cháy rừng phục vụ phòng chống, giảm thiểu thiệt hại do cháy rừng tại tỉnh Sơn La, Việt Nam. *Tạp chí Khoa học ĐHQGHN: Các Khoa học Trái đất và Môi trường*, 33(3), 53-66.

[5] Taylor, S. W., Woolford, D. G., Dean, C. B. & Martell, D. L., 2013. Wildfire Prediction to Inform Fire Management: Statistical Science Challenges. *Statist. Sci.*, 28(4), 586-615.

DEVELOPING A DEEP NEURAL NETWORK MODEL FOR PREDICTING FOREST FIRE RISK OF LAM DONG PROVINCE

Le Van Hung[1], Nguyen Thi Thanh[1], Dang Huu Nghi[1], Hoang Anh Duc[1]

[1] *Hanoi University of Mining and Geology*, {levanhung@humg.edu.vn, nguyenthithanh@humg.edu.vn, danghuunghi@humg.edu.vn, hoanganhduc@humg.edu.vn}

ABSTRACT

Forest fires can cause serious damage to property and life of humans, plants, animals, ecosystems, and the environment. Due to climate change as well as human activities, forest fires have risen to alarming levels in Vietnam. Forest fire susceptibility prediction is an important task in forest fire prevention and control. In this paper, we develop a deep feedforward neural network model for predicting and producing a forest fire susceptibility map of Lam Dong province. Input factors consist of slope, aspect, elevation, land use, NDVI, distance to roads, distance to residence areas, temperature, wind speed, and rainfall. The results show that the model performs well on both the training dataset and the validation dataset. The performance of the model is also compared with that of several other machine learning models. The built model is then used to compute the forest fire susceptibility indexes and create a forest fire susceptibility map for the study area.

Key words: Deep Feedforward Network, Machine Learning, Forest Fire, Lam Dong province