Ứng dụng các phương pháp học tập kết hợp trong dự báo nguy cơ cháy rừng tại Gia Lai

ĐẶNG HỮU NGHỊ; BÙI THỊ VÂN ANH
Trường Đại học Mỏ - Địa chất Hà Nội

So với cả vùng Tây Nguyên, Gia Lai chiếm 28% diện tích lâm nghiệp, 30% diện tích có rừng và 38% trữ lượng gỗ. Nằm trong vùng có điều kiện khí hậu, địa hình, đất đai nhiều thuận lợi, nên thảm thực vật ở đây phát triển rất đa dạng và phong phú, bao gồm nhiều loại khác nhau. Rừng tự nhiên ở Gia Lai chiếm khoảng 78,3% diện tích đất lâm nghiệp, có nhiều loại cây quý hiếm, gỗ tốt. Cháy rừng là mối đe dọa lớn, ảnh hưởng xấu đến môi trường và vùng sinh thái. Do đó, theo dõi hiện trạng và dự báo cháy rừng là rất cần thiết nhằm góp phần bảo vệ tài nguyên rừng. Tổng quan các công trình nghiên cứu cho thấy hiện nay trên thế giới vẫn chưa có phương pháp chung cho bài toán dự báo nguy cơ cháy rừng. Trong bài báo này, chúng tôi thử nghiệm áp dụng và so sánh các phương pháp học tập thể (Ensemble Learning) cho bài toán dự báo nguy cơ cháy rừng tại Gia Lai.

Abstract: *Compared with the whole Central Highlands, Gia Lai accounts for 28% of the forestry area, 30% of the forested area and 38% of the timber reserves. Located in an area with favorable climate, terrain and soil conditions, the vegetation here is very diverse and rich, including many different types. Natural forests in Gia Lai account for about 78.3% of the forestry land area, with many rare trees and good timbers. Forest fires are a great threat, adversely affecting the environment and ecological regions. Therefore, monitoring the current status and forecasting of forest fires is very necessary to contribute to the protection of forest resources. An overview of research studies shows that currently in the world there is still no general method for the problem of predicting the risk of forest fires. In this paper, we try to apply and compare Ensemble Learning methods to the problem of forest fire risk prediction in Gia Lai.*

Giới thiệu

Các phương pháp thống kê cũng được sử dụng để mô hình hóa cháy rừng do tính chất ngẫu nhiên cố hữu của hiện tượng cháy rừng ở tất cả các quy mô. Do cháy rừng một quá trình phức tạp, nên trong các bài toán mô hình hóa cháy rừng với nhiều yếu tố ảnh hưởng và khối lượng dữ liệu lớn (cho vùng nghiên cứu rộng), độ chính xác dự báo của các mô hình thống kê vẫn còn hạn chế.

Để nâng cao độ chính xác dự báo của các mô hình cháy rừng, các kỹ thuật học máy đã được đề xuất do chúng làm việc tốt với dữ liệu lớn, có nhiều đầu vào. Trong thống kê và học máy, các phương pháp học tập kết hợp sử dụng nhiều thuật toán học tập để có được hiệu suất dự đoán tốt hơn những gì có thể thu được từ bất kỳ thuật toán học tập cấu thành nào. Nhiều nghiên cứu thực nghiệm và lý thuyết đã chứng minh rằng các mô hình kết hợp thường đạt độ chính xác cao hơn các mô hình đơn lẻ.

Phương pháp học tập kết hợp

Phương pháp học tập kết hợp là kỹ thuật tạo ra nhiều mô hình và sau đó kết hợp chúng lại để tạo ra kết quả được cải thiện hơn. Các phương pháp Ensemble Learning được chia thành 3 loại: Bagging (đóng bao), Boosting (tăng cường) và Stacking (Xếp chồng).

Bagging: Thuật toán Bagging được đề xuất bởi Breiman [4] và mục đích của nó là để cải thiện hiệu quả dự đoán đối với vấn đề mất cân bằng dữ liệu khi chỉ áp dụng một thuật toán đơn như Decision tree hoặc Neural Network.

Boosting: Boosting được giới thiệu bởi [9] sử dụng thuật toán cây quyết định để tạo các mô hình mới. Boosting gán trọng số cho các mô hình dựa trên hiệu suất của chúng. Có nhiều biến thể của thuật toán Boosting như LogitBoost (LB) và AdaBoost (AB).

Stacking: Stacking là một biến thể của mô hình máy học kết hợp - ensemble learning còn được gọi là phương pháp meta-learning, bao gồm một hệ thống phân cấp các bộ phân loại khác nhau. Mục tiêu của stacking là để xây dựng một bộ phân loại cấp độ meta có thể dự đoán nhãn đích của tập dữ liệu bằng cách kết hợp kết quả các dự đoán từ các bộ phân loại riêng biệt.

Thực nghiệm

Tập dữ liệu

Cơ sở dữ liệu GIS của bài toán dự báo nguy cơ cháy rừng cho tỉnh Gia Lai bao gồm dữ liệu của 12 yếu tố ảnh hưởng và 2530 vị trí cháy rừng trong giai đoạn năm 2007-2016. Các yếu tố này bao gồm: Độ dốc địa hình, hướng phơi sườn, độ cao, độ cong địa hình, hiện