

Ứng dụng trí tuệ nhân tạo trong phân loại cây trồng

○ BÙI THỊ VÂN ANH¹; ĐẶNG HỮU NGHỊ¹

¹ Trường Đại học Mở - Địa chất

Tóm tắt

Việc phân loại cây trồng chính xác đóng vai trò quan trọng trong quản lý nông nghiệp, dự báo sản lượng và giám sát an ninh lương thực. Trong những năm gần đây, sự phát triển mạnh mẽ của trí tuệ nhân tạo (AI), đặc biệt là học sâu (deep learning) đã mở ra cơ hội mới để khai thác dữ liệu quan sát Trái đất nhằm phân loại cây trồng trên diện rộng. Bài báo này trình bày nghiên cứu ứng dụng AI cho bài toán phân loại cây trồng dựa trên bộ dữ liệu CropHarvest, một tập dữ liệu mở quy mô toàn cầu về loại cây trồng, thu thập từ nhiều nguồn viễn thám. Nghiên cứu sử dụng các mô hình học máy và học sâu, bao gồm Random Forest, XGBoost và mạng nơ-ron tích chập (CNN) trên chuỗi thời gian, để huấn luyện và đánh giá hiệu quả phân loại. Kết quả cho thấy, mô hình học sâu dựa trên dữ liệu chuỗi thời gian đạt độ chính xác cao hơn, đặc biệt đối với các loại cây có tính mùa vụ rõ rệt.

Từ khóa: Trí tuệ nhân tạo, phân loại cây trồng, CropHarvest, học sâu, viễn thám.

Đặt vấn đề

Phân loại cây trồng là một trong những nhiệm vụ cốt lõi trong nông nghiệp số, giúp các nhà quản lý nông nghiệp và Chính phủ: Dự báo năng suất; quản lý tài nguyên đất đai; đánh giá tác động của biến đổi khí hậu và hỗ trợ lập chính sách an ninh lương thực.

Các phương pháp truyền thống dựa vào khảo sát thực địa thường tốn kém và không hiệu quả trên diện rộng. Trong bối cảnh đó, dữ liệu viễn thám kết hợp với trí tuệ nhân tạo đã chứng minh khả năng vượt trội trong việc nhận diện, phân loại và giám sát cây trồng toàn cầu.

Bộ dữ liệu CropHarvest (Tseng, G., et al., 2021)^[1] là một trong những tập dữ liệu lớn và toàn diện nhất hiện nay, bao gồm nhãn cây trồng ở nhiều khu vực trên thế giới, kết hợp với ảnh vệ tinh từ Sentinel và MODIS. Bộ dữ liệu mở ra cơ hội nghiên cứu ứng dụng AI trong phân loại cây trồng trên phạm vi toàn cầu.

Cơ sở lý thuyết

Trí tuệ nhân tạo trong nông nghiệp

AI ứng dụng trong nông nghiệp chủ yếu tập trung vào ba lĩnh vực: Nhận diện và phân loại cây trồng (crop classification); dự báo năng suất (yield prediction); phát hiện dịch bệnh và stress môi trường (disease and stress detection). Trong đó, phân loại cây trồng dựa trên dữ

liệu viễn thám được xem là bài toán nền tảng.

Bộ dữ liệu CropHarvest

Phạm vi: Toàn cầu, bao gồm nhiều châu lục.

Nội dung: Gồm nhãn cây trồng (crop type), ảnh vệ tinh (Sentinel-2, MODIS), dữ liệu khí hậu.

Kích thước: Hơn 90.000 mẫu với nhiều loại cây như lúa, ngô, đậu nành, lúa mì,...



Định dạng: Chuỗi thời gian của phổ phản xạ vệ tinh, cùng với thông tin không gian (tọa độ).

Phương pháp phân loại

Phương pháp Rừng ngẫu nhiên (Random Forest - RF): Rừng ngẫu nhiên gồm một tổ hợp các cây quyết định không cắt nhánh. Mỗi cây quyết định được xây dựng bởi thuật toán CART trên tập mẫu lấy mẫu ngẫu nhiên có hoàn lại từ tập dữ liệu ban đầu. Tại mỗi nút, một phân hoạch tốt nhất được thực hiện dựa trên thông tin trong một không gian con các thuộc tính được chọn ngẫu nhiên từ không gian thuộc tính ban đầu. RF tổng hợp kết quả dự đoán của các cây quyết định làm kết quả cuối cùng.

Phương pháp XGBoost: Là một thư viện học máy mã nguồn mở, hiệu quả và phổ biến, nổi bật với tốc độ, khả năng xử lý dữ liệu lớn và tính linh hoạt. Nó là một thuật toán tăng cường gradient cực đoan, được sử dụng rộng rãi cho các bài toán phân loại, hồi quy và xếp hạng, đặc biệt là trong các cuộc thi về khoa học dữ liệu. XGBoost được xây dựng dựa trên nền tảng của gradient boosting nhưng có nhiều cải tiến như chính quy hóa để tránh hiện tượng quá khớp, xử lý song song, và khả năng xử lý giá trị thiếu một cách tự động.

Phương pháp Mạng nơ-ron tích chập (CNN) trên chuỗi thời gian: Mô hình CNN cho dữ liệu chuỗi thời gian là một loại kiến trúc mạng nơ-ron thường được sử dụng cho các tác vụ liên quan đến dữ liệu tuần tự, chẳng hạn như dự báo chuỗi thời gian hoặc phát hiện bất thường. CNN, vốn thường được sử dụng cho các tác vụ xử lý hình ảnh, đã được điều chỉnh để xử lý dữ liệu tuần tự bằng cách áp dụng tích chập trên toàn bộ chiều thời gian.

Phương pháp LSTM (Long Short-Term Memory): Là kiến trúc mạng nơ-ron hồi quy (Recurrent Neural Networks - RNN) được thiết kế nhằm khắc phục hạn chế trong việc ghi nhớ thông tin dài hạn. Trong khi RNN chỉ sử dụng một trạng thái ẩn đơn giản để truyền thông tin qua các bước thời gian, thì LSTM bổ sung một bộ nhớ tế bào (memory cell) cùng với ba cơ chế cổng kiểm soát: Cổng quên, cổng đầu vào và cổng đầu ra. Các cổng này cho phép mô hình lựa chọn thông tin cần ghi nhớ hoặc loại bỏ, nhờ đó duy trì được ngữ cảnh trong các chuỗi dữ liệu dài.

Phương pháp Transformer: Là kiến trúc mô hình học sâu dựa trên cơ chế self-attention, cho phép mô hình này hiểu được mối quan hệ giữa các từ trong một câu mà không cần đến kiến trúc tuần tự truyền thống như RNN (Recurrent Neural Networks) hay LSTM (Long Short-Term Memory). Transformer có khả năng xử lý toàn bộ câu cùng một lúc, điều này giúp tăng tốc độ huấn luyện và cải thiện hiệu quả xử lý.

Phương pháp nghiên cứu

Quy trình nghiên cứu: Tiền xử lý dữ liệu: Lọc dữ liệu thiếu, chuẩn hóa giá trị phổ vệ tinh. Chia tập dữ liệu: 70% huấn luyện, 15% kiểm định, 15% kiểm tra. Xây dựng mô hình: Random Forest (baseline); XGBoost; CNN 1D cho chuỗi thời gian Sentinel-2.

Đánh giá mô hình: Accuracy, F1-score, ma trận nhầm lẫn.

Kiến trúc CNN 1D: Input: Chuỗi thời gian 12 tháng dữ liệu NDVI/EVI. Lớp convolution (kernel size=3, stride=1). Lớp pooling, dropout chống overfitting. Fully-connected + softmax cho phân loại đa lớp.

Thực nghiệm và kết quả

Random Forest: Accuracy = 78%, F1 = 0.75. XGBoost: Accuracy = 82%, F1 = 0.80. CNN (1D): Accuracy = 88%, F1 = 0.86.

Ma trận nhầm lẫn cho thấy, các loại cây ngắn ngày (ví dụ: Đậu nành, khoai) dễ bị nhầm lẫn, trong khi lúa và ngô được phân loại chính xác cao hơn.

Kết luận

Bài báo đã trình bày ứng dụng AI trong phân loại cây trồng dựa trên bộ dữ liệu CropHarvest. Kết quả cho thấy, các mô hình học sâu, đặc biệt là CNN 1D, vượt trội so với các thuật toán học máy truyền thống. Hướng nghiên cứu tương lai bao gồm: Sử dụng mô hình Transformer và attention-based. Kết hợp thêm dữ liệu khí hậu và thổ nhưỡng. Triển khai trên phạm vi cục bộ để hỗ trợ quản lý nông nghiệp thông minh.

Tài liệu tham khảo

1. Tseng, G., *et al.* (2021). CropHarvest: A global dataset for crop-type classification. *Scientific Data*;
2. Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32;
3. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of KDD*;
4. Russakovsky, O., *et al.* (2015). ImageNet Large Scale Visual Recognition Challenge. *IJCV*. ■