ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA: CÔNG NGHỆ THÔNG TIN

NHẬN DẠNG

HW01: Nhận dạng mẫu

Giảng viên hướng dẫn: Thầy Lê Hoàng Thái, Thầy Lê Thanh Phong
Họ tên: Phạm Dương Trường Đức
MSSV: 20120061

Nội dung

I. TỔNG QUAN

1. Giới thiệu bài toán

Trong nông nghiệp, năng suất của cây trồng phụ thuộc rất nhiều vào thời tiết, chế độ chăm sóc, sâu bệnh,… để tăng năng suất người trồng cần có những biện pháp hạn chế những tác động của các yếu tố trên.

Tuy nhiên, bản thân cây trồng cũng có những loại bệnh nhất định và thường biểu hiện ra bên ngoài trên thân, lá… những biểu hiện đó bất thường trên cây trồng thường liện quan tới một loại bệnh nào đó.

Nhằm giúp người trồng xác định được loại bệnh đang có trên cây trồng, với tập dữ liệu hình ảnh về các loại bệnh trên cây, ta sẽ phân tích và xây dựng mô hình phân lớp bệnh trên lá cây bao gồm 4 loại: Combinations, Healthy, Rust và Scab.

2. Phân tích bài toán

Đặc điểm các loại bệnh:

• Rust: là bệnh xuất hiện những đốm màu vàng hoặc vàng cam trên lá cây.

• Scab: là bệnh xuất hiện những vệt màu nâu xuất hiện khắp lá cây.

• Multiply_disease: là bệnh xuất hiện cả những đốm vàng và vệt nâu trên lá cây.

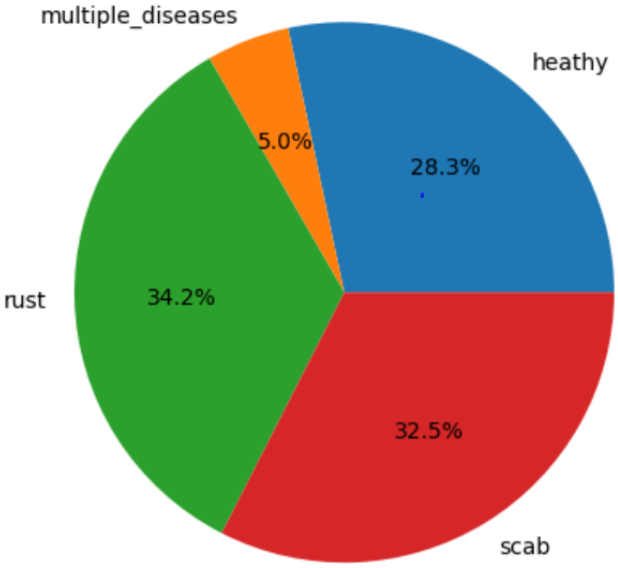

Figure 1: Biểu đồ tròn thể hiện số lượng mỗi loại bệnh

Có thể thấy số lượng mỗi loại bệnh khá tương đồng. Tuy nhiên, số lượng bệnh hỗn hợp trên lá ít hơn đáng kể so với so với các bệnh còn lại. Điều này có thể dẫn đến sự cố mất cân bằng trong quá trình đào tạo đối với loại bệnh này.

Đối với một bài toán nhận dạng mẫu, chúng ta sẽ thực hiện các bước sau:

- Chuẩn bị tập dữ liệu huấn luyện và tập dữ liệu test (đã được cho).
- Rút trích đặc trưng bộ hình ảnh
- Sử dụng một thuật toán học máy để tiến hành nhận dạng.

II. Phương pháp thực hiện

1. Rút trích đặc trưng

Ở bài toán này, em sử dụng mô hình **VGG16** để rút trích đặc trưng.

Mô hình VGG16 là một mô hình mạng tích chập nơ-ron (CNN), được huấn luyện trên bộ dữ liệu ImageNet, gồm hơn 1 triệu ảnh và được phân loại vào 1000 lớp khác nhau. Với kiến trúc đơn giản, VGG16 có khả năng rút trích đặc trưng từ ảnh rất hiệu quả. Mô hình VGG16 sử dụng các lớp tích chập với kernel size là 3x3 và các lớp pooling để giảm kích thước của feature map. Cuối cùng, VGG16 kết thúc với một số lớp fully-connected để phân loại đối tượng trên ảnh.

VGG16 đã trở thành một trong những mô hình rất được ưa chuộng để rút trích đặc trưng trong các bài toán liên quan đến xử lý ảnh, như phân loại, phát hiện đối tượng, phân tích bức ảnh.

2. Huấn luyện mô hình

Sau khi rút trích đặc trưng, em sẽ thực hiện huấn luyện mô hình SVM. Dựa trên bộ hình ảnh train, em thực hiện chia bộ dữ liệu đã được rút trích đặc trưng và nhãn của chúng thành tập train và tập test với tỷ lệ 80/20 để đưa vào mô hình SVM.

SVM có khả năng phân loại tốt cho các bộ dữ liệu lớn, đặc biệt là trong bài toán phân loại hình ảnh. Trong việc phân loại hình ảnh, SVM thường được sử dụng để phân loại các đối tượng trong hình ảnh.

Thuật toán này hoạt động bằng cách tìm một ranh giới phân cách tốt nhất giữa các lớp đối tượng khác nhau. SVM sử dụng một hàm kernel để chuyển đổi các đặc trưng của ảnh sang một không gian mới, từ đó tạo ra một ranh giới phân cách tốt nhất giữa các lớp đối tượng.

Trong quá trình huấn luyện, SVM cần được cung cấp các đặc trưng của ảnh để học cách phân loại các đối tượng khác nhau. Sau khi được huấn luyện, SVM sẽ dùng để dự đoán lớp của các đối tượng trong ảnh mới.

3. Kết quả và đánh giá

```
Accuracy test: 0.7972602739726027
              precision    recall  f1-score   support

           0       0.77      0.83      0.80       100
           1       1.00      0.00      0.00        18
           2       0.82      0.86      0.84       120
           3       0.80      0.83      0.81       127

    accuracy                           0.80       365
   macro avg       0.85      0.63      0.61       365
weighted avg       0.81      0.80      0.78       365
```

Có thể thấy mô hình hoạt động tương đối tốt, tuy nhiên nhãn ‘multiple-diseases’ nhận diện kém do thiếu dữ liệu.

Để giải quyết vấn đề này, có thể sử dụng các kỹ thuật như lấy mẫu quá mức lớp thiểu số, lấy mẫu dưới lớp đa số hoặc sử dụng các thuật toán phức tạp hơn như phương pháp tập hợp để cải thiện hiệu suất của mô hình.

III. Kết luận

Hiện nay, bài toán nhận diện, phân loại hình ảnh đang trở nên ngày càng phổ biến và quan trọng trong nhiều lĩnh vực. Trong đó, bài toán nhận diện bệnh trên lá cây đóng vai trò quan trọng trong việc giám sát và quản lý sức khỏe cây trồng, góp phần tối ưu hóa sản xuất nông nghiệp và bảo vệ môi trường. Để xây dựng mô hình nhận diện, phân loại tốt, chúng ta cần phải có kiến thức về các loại bệnh để phân tích và tìm ra những đặc trưng cần thiết để phân biệt chúng. Đồng thời, kĩ năng phân tích dữ liệu cũng rất quan trọng để chọn ra các thuật toán phù hợp và có kết quả tốt nhất. Việc tiền xử lý dữ liệu như giảm độ nhiễu, gia tăng số lượng ảnh, cũng như kết hợp các mô hình deep learning như CNN sẽ giúp cho mô hình nhận diện, phân loại bệnh trên lá cây đạt được kết quả dự đoán chính xác và tin cậy.

Tài liệu tham khảo

1. Plant disease Identification using SVM: https://ijcrt.org/papers/IJCRT2203059.pdf
2. Transfer learning using CNN (VGG16) as feature extractor and Random Forest classifier: https://www.youtube.com/watch?v=IuoEiemAuIY
3. Nhận dạng mẫu và ứng dụng thử nghiệm - Trí tuệ nhân tạo - Nhận dạng: https://www.youtube.com/watch?v=Pfd9vGIlB9c&list=PLFpf6IAKcCjVwoHYSGENBdNLkQVvubEsX&index=2