

Báo cáo nhóm đồ án cuối kì môn Machine Learning

PHẦN 1: THÔNG TIN TÓM TẮT

Tên đề tài (IN HOA)	PHÂN LOẠI TRÁI CÂY DỰA TRÊN HÌNH ẢNH		
Họ và tên (IN HOA)	VÕ HUY KHÔI	LÊ ĐOÀN NHẬT MINH	NGUYỄN THỊNH QUYỀN
Lớp - MSSV	18520949	18521101	18521322
Ảnh			
Link Github chứa repos CS114.K21	https://github.com/HuyKhoi-code/CS114.K21	https://github.com/nhatminh-it/CS114.K21	https://github.com/dangcaptopkd/CS114.K21
	Link đồ án: https://github.com/nhatminh-it/fruits_classification		
Điểm đánh giá giữa kỳ (A B C D)	C	B	B
Thành tích để tính điểm bonus			

	<ul style="list-style-type: none"> - Đồ án có app để test: https://github.com/nhatminh-it/fruits_classification/blob/master/Fruits_classify_app.ipynb - Cả 3 đều có chứng chỉ Intro to Machine Learning của Kaggle 		
Tóm tắt Bài tập quá trình	Số lần nộp bài tập Quá trình trên Classroom: 36/36 Số lần nộp bài Thực hành trên Classroom: 08/08 Tự đánh giá: 90/100	Số lần nộp bài tập Quá trình trên Classroom: 36/36 Số lần nộp bài Thực hành trên Classroom: 08/08 Tự đánh giá: 100/100	Số lần nộp bài tập Quá trình trên Classroom: 36/36 Số lần nộp bài Thực hành trên Classroom: 08/08 Tự đánh giá: 100/100
Tóm tắt Đồ án Cuối kỳ (không quá 500 từ)	<ul style="list-style-type: none"> - Sử dụng Machine Learning để xây dựng model phân loại trái cây dựa trên hình ảnh. Từ đó, có thể giúp quá trình sắp xếp và phân loại trái cây trở thành một quá trình tự động gần như hoàn toàn. <ul style="list-style-type: none"> ● Input: Ảnh của một loại trái cây trên nền đơn giản (1 màu). ● Output: Tên của loại trái cây đó. - Cách thức giải quyết: Sử dụng các phương pháp xử lý ảnh để trích xuất đặc trưng sau đó dùng model để train - Kết quả: Kết quả đạt được tỉ lệ chính xác trên tập train và test ~ 100% - Tự đánh giá: 95/100 		
Link khác			

PHẦN 2: BÁO CÁO TÓM TẮT ĐỒ ÁN CUỐI KỲ

I. Tổng quan bài toán

Phân loại trái cây là một lĩnh vực nghiên cứu vô cùng hứa hẹn với sự đa dạng của chủng loại trái cây. Nghiên cứu này đặc biệt vô cùng hữu ích đối với việc giám sát và phân loại các loại trái cây để đảm bảo cho chuỗi cung cấp thực phẩm diễn ra nhanh hơn. Có thể áp dụng nghiên cứu cho việc sắp xếp hàng hóa diễn ra tự động hoặc nhận diện trái cây để tính tiền mà không cần quét mã vạch. Trong đề tài này, nhóm chỉ tập trung đặc biệt đến sự phân loại trái cây ở các chuỗi cung cấp như siêu thị hoặc tạp hóa. Để tăng tính thực tế cho bài toán, nhóm chỉ phân loại 10 loại trái cây nằm trong top 12 loại trái cây được tiêu thụ nhiều nhất thế giới: cà chua, chuối, dưa hấu, táo, bưởi chùm (nhóm không kiểm được), nho, cam, xoài, chuối mề (nhóm không kiểm được), quýt, lê, dứa.

II. Hướng giải quyết

- Input: một hình ảnh trái cây nằm trong top 10 loại trái được tiêu thụ nhiều nhất với phong nền sáng.
- Output: tên của loại trái đó.

Sử dụng các mô hình thuật toán Machine learning như Support Vector Machine và Random Forest cùng với các thuật toán xử lý ảnh để trích xuất đặc trưng của trái cây từ 1 tấm ảnh. Các đặc trưng của trái cây được nhóm sử dụng là màu sắc chính của trái và mật độ cạnh của trái trong ảnh.

III. Mô tả dữ liệu

2500 tấm ảnh tự chụp của các 10 loại quả mà nhóm có thể kiếm được tại siêu thị Big C Dĩ An và được resize lại thành 200 x 200 pixels. Mỗi loại trái: 250 tấm ảnh.

Bộ dữ liệu này được chia thành 2 tập train và valid:

- Train: 187 ảnh

- Valid: 63 ảnh

Bộ data test model (hay còn được lưu thành bộ valid data trong github) thực tế với các ảnh được download trên google hình ảnh: 169 ảnh.

IV. Các hướng xử lý ảnh

Resize ảnh về kích thước 200 x 200 pixels để giảm thời gian load ảnh

- Trích xuất đặc trưng màu:

- Trích xuất đặc trưng cạnh của trái cây: sử dụng thuật toán Canny edge của OpenCV 2 để đưa ảnh về các cạnh đặc trưng và sử dụng các vùng 40 x 40 pixels để tính mean và standard deviation (std) của từng vùng.

V. Kết quả đánh giá

Thực hiện các model machine learning đối với 4 hướng tiếp cận:

- Sử dụng trung bình và độ lệch chuẩn của các giá trị pixels cho từng lớp RGB của ảnh đã cluster:

- Cluster 1 (50 columns)

RandomForestClassifier

- Train accuracy: 1.0
- Test accuracy: 0.8365079365079365

SVC

- Train accuracy: 0.9096256684491979
- Test accuracy: 0.8269841269841269

- Sử dụng trung bình và độ lệch chuẩn của các giá trị pixels cho toàn ảnh đã được cluster về 2 màu với các section 40 x 40 pixels:

- Cluster 2 (6 columns)

RandomForestClassifier

- Train accuracy: 1.0
- Test accuracy: 0.9142857142857143

SVC

- Train accuracy: 0.8978609625668449
- Test accuracy: 0.9047619047619048

- Sử dụng trung bình và độ lệch chuẩn của các giá trị pixels cho từng section 40 x 40 với ảnh gốc được xử lý về canny edge:

- random forest classify:

test accuracy: 0.7746031746031746

- svc:

test accuracy: 0.8047619047619048

- Sử dụng trung bình và độ lệch chuẩn của các giá trị pixels cho từng section 40 x 40 với ảnh đã được cluster chỉ còn 2 màu và xử lý về canny edge:

- random forest classify:
test accuraccy: 0.7142857142857143
- SVC:
test accuraccy: 0.6984126984126984

VI. Kết luận

Sử dụng kết hợp 2 phương pháp cluster đưa ảnh về còn 2 màu chính trích xuất đặc trưng là mean và std của 3 lớp RGB màu chính và sử dụng canny edge để trích xuất đặc trưng mean và std của các section.

- RandomForestClassifier
 - best param: {'max_depth': 15, 'n_estimators': 300}
 - train accuracy: 1.0
 - test accuracy: 0.9523809523809523
 - Những trái đúng 100% trong tập test: cà chua, cam, nho, thơm
- SVC
 - best params: {'degree': 3, 'kernel': 'linear'}
 - train accuracy: 1.0
 - test accuracy: 0.9412698412698413
 - Những trái đúng 100% trong tập test: dưa hấu

Kết quả thực nghiệm của model đối với tập test data:

```

tao
15/16 wrong
nho
8/12 wrong
quyt
3/3 wrong
xoai
19/19 wrong
thom
13/19 wrong
cam
3/19 wrong
chuoi
16/22 wrong
le
20/20 wrong
duahau
16/19 wrong
cachua
2/20 wrong

```