TRƯỜNG ĐẠI HỌC CẦN THƠ

TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

ĐỒ ÁN HỌC PHẦN

NGUYÊN LÝ MÁY HỌC

Đề tài:

ĐÁNH GIÁ CÁC MÔ HÌNH MÁY HỌC TRÊN TẬP DỮ LIỆU VĂN BẢN, HÌNH ẢNH VÀ KẾT HỢP VĂN BẢN VỚI HÌNH ẢNH

Giảng viên hướng dẫn:

TS.Phạm Thế Phi

Sinh viên thực hiện :

Huỳnh Minh Luân B2106842

Cần Thơ, 11/2024

NHẬN XÉT CỦA GIẢNG VIÊN

Cần Thơ, ngày tháng năm

(Ký và ghi rõ họ tên)

LỜI CẢM ƠN

Để hoàn thành bài đồ án học phần này, em xin gửi lời cảm ơn chân thành và sâu sắc đến thầy Phạm Thế Phi – người đã tận tâm hướng dẫn và giúp đỡ em trong suốt quá trình thực hiện báo cáo. Nhờ vào sự chỉ dẫn quý báu của thầy, em đã hoàn thành bài báo cáo một cách tốt nhất.

Em cũng xin gửi lời cảm ơn sâu sắc đến các Thầy Cô tại Trường Đại học Cần Thơ, đặc biệt là các Thầy Cô ở Trường CNTT & TT, những người đã truyền đạt những kiến thức quý giá trong suốt thời gian qua.

Bên cạnh đó, em xin chân thành cảm ơn bạn bè và gia đình đã luôn động viên, khích lệ và tạo điều kiện giúp đỡ em, góp phần quan trọng giúp em hoàn thành bài báo cáo này.

Mặc dù em đã cố gắng hết sức trong quá trình thực hiện niên luận cơ sở, nhưng không thể tránh khỏi những thiếu sót. Em rất mong nhận được những ý kiến đóng góp quý báu từ quý Thầy Cô và các bạn để bài báo cáo được hoàn thiện hơn.

Cần Thơ, ngày 06 tháng 11 năm 2024

Người viết

Huỳnh Minh Luân

MỤC LỤC

CHƯƠNG 1 GIỚI THIỆU

1.1 GIỚI THIỆU ĐỀ TÀI

1.1.1 Giới thiệu tổng quan về đề tài môn học

Đồ án đánh giá hiệu quả của các mô hình máy học dựa trên những tập dữ liệu hình ảnh, tập dữ liệu văn bản và kết hợp để cải thiện độ chính xác trong việc phần loại.

Các mô hình máy học được sử dụng trong đồ án là các mô hình máy học phổ biến như **Nearest Neighbors, Linear SVM, RBF SVM, Decision Tree, Random Forest, Neural Net, AdaBoot, Naïve Bayes.** Mục tiêu của đồ án nhằm đánh giá hiệu xuất của các mô hình trên các tập dữ liệu khác nhau để chọn ra các mô hình thích hợp nhất với các loại dữ liệu tương ứng.

1.1.2 Yêu cầu đề tài

- Chạy và so sánh hiệu quả phân lớp của các mô hình máy học trên dữ liệu văn bản. Cho nhận xét và giải thích trên mô hình máy học cho độ chính xác cao nhất.
- Chạy và so sánh hiệu quả phân lớp của các mô hình máy học trên dữ liệu ảnh. Cho nhận xét và giải thích trên mô hình máy học cho độ chính xác cao nhất.
- Chạy và so sánh hiệu quả phân lớp của các mô hình máy học kết hợp trên dữ liệu văn bản và hình ảnh. Hãy chỉ ra phương pháp kết hợp phân loại dữ liệu và hình ảnh cho độ chính xác phân lớp cao nhất và giải thích.

1.2 GIỚI THIỆU HỆ THỐNG SỬ DỤNG

Sử Dụng Gói phần mềm mlfw sử dụng ngôn ngữ Python, bao gồm các modules sau đây :

- **Helpers** : Dùng để tiền xử lý dữ liệu văn bản hoặc hình ảnh, rút trích đặc trưng dữ liệu văn bản hoặc hình ảnh. Đặc trưng văn bản gồm TF-ID. Đặc trưng hình ảnh gồm: màu sắc (color), SIFT, hog, gist, deep (có thể chọn vgg16, vgg19,inception, xception, resnet).

– **Processors** : Dùng để chuẩn bị, nạp dữ liệu (Data), huấn luyện và thử nghiệm các mô hình máy học trên mô thức dữ liệu văn bản (Text) và hình ảnh (Image). Module Processors sử dụng các hàm tiền xử lý dữ liệu và rút trích đặc trưngdữ liệu từ molule Helpers.

– **Ranker** : Dùng để gọi chạy các thành phần của module Processors.

– **Runner** : Gọi chạy Ranker, truyền 03 tham số: Đường dẫn đến thư mục chứa tập dữ liệu huấn luyện, thử nghiệm và số lớp.

1.3 GIỚI THIỆU TẬP DỮ LIỆU

Tập dữ liệu bao gồm hình ảnh và văn bản để mô tả 2 lớp dữ liệu bao gồm Giầy và Túi Xách bao gồm dữ liệu dạng văn bản và dữ liệu dạng hình ảnh.

	Túi xách		Giày	
	Train	Test	Train	Test
Văn bản	18	20	18	20
Hình ảnh	18	20	18	20

1.4 BỐ CỤC BÁO CÁO

Chương 1: Giới Thiệu

- Giới thiệu về đề tài thực hiện , hệ thống sử dụng để đánh giá và giới thiệu về tập dữ liệu sử dụng

Chương 2: Nội Dung

- Các mô hình với tập dữ liệu Văn Bản , Hình Ảnh và Kết Hợp

Chương 3: Kết Luận

- Kết quả đạt được sau nghiên cứu và hướng phát triển

CHƯƠNG 2 NỘI DUNG

2.1 CÁC MÔ HÌNH MÁY HỌC DÙNG TRÊN TẬP VĂN BẢN

2.1.1 Kết quả chạy thử nghiệm

Mô hình	Độ chính xác	Thời gian xử lý
Nearest Neighbors	0.95	0.007 giây
Linear SVM	0.875	1.295 giây
RBF SVM	0.8	1.151 giây
Decision Tree	0.65	0.002 giây
Random Forest	0.725	0.063 giây
Neural Net	0.85	0.288 giây
AdaBoost	0.85	0.147 giây
Naive Bayes	0.975	0.001 giây

So sánh chi tiết các mô hình

- **Naive Bayes và Nearest Neighbors:** Hai mô hình này đạt được độ chính xác cao nhất (0.975). Điều này cho thấy sự phù hợp của các mô hình này với bài toán phân lớp văn bản.
- **Linear SVM, RBF SVM, Neural Net và AdaBoost:** Các mô hình này cũng đạt được độ chính xác khá cao (từ 0.8 đến 0.85).
- **Decision Tree và Random Forest:** Mặc dù thời gian xử lý nhanh nhưng độ chính xác của hai mô hình này lại thấp hơn so với các mô hình còn lại.

Nhận xét về mô hình có độ chính xác cao nhất (Naive Bayes và Nearest Neighbors)

- **Naive Bayes:**
  - **Ưu điểm:** Đạt được độ chính xác cao, đơn giản và hiệu quả tính toán, phù hợp với dữ liệu văn bản có nhiều thuộc tính.
  - **Nhược điểm:** Dựa trên giả định về độc lập giữa các thuộc tính, có thể không chính xác trong một số trường hợp.
- **Nearest Neighbors:**
  - **Ưu điểm:** Đơn giản, không cần huấn luyện mô hình, phù hợp với các bài toán có ít lớp.
  - **Nhược điểm:** Độ phức tạp tính toán tăng theo số lượng dữ liệu huấn luyện, dễ bị ảnh hưởng bởi dữ liệu nhiễu.

Giải thích tại sao Naive Bayes và Nearest Neighbors đạt được độ chính xác cao

- **Naive Bayes:** Với dữ liệu văn bản, ta thường biểu diễn văn bản dưới dạng một vector các từ (bag-of-words). Naive Bayes sử dụng định lý Bayes để tính xác suất một văn bản thuộc

một lớp nhất định dựa trên sự xuất hiện của các từ trong văn bản đó. Mặc dù giả định về độc lập giữa các từ không hoàn toàn đúng, nhưng Naive Bayes vẫn hoạt động khá tốt trong thực tế.

- **Nearest Neighbors:** Mô hình này phân loại một văn bản mới dựa trên các văn bản gần nhất trong tập dữ liệu huấn luyện. Với dữ liệu văn bản, ta thường sử dụng các phương pháp đo lường khoảng cách như cosine similarity để tính toán sự tương tự giữa các văn bản.

2.2 CÁC MÔ HÌNH MÁY HỌC DÙNG TRÊN TẬP HÌNH ẢNH

2.2.1 Kết quả chạy thử nghiệm

Mô hình	Độ chính xác	Thời gian xử lý
Nearest Neighbors	0.65	0.135 giây
Linear SVM	0.975	4.053 giây
RBF SVM	0.525	38.978 giây
Decision Tree	0.9	0.134 giây
Random Forest	0.925	0.047 giây
Neural Net	0.975	6.189 giây
AdaBoost	0.9	0.166 giây
Naive Bayes	0.95	0.163 giây

So sánh chi tiết các mô hình

- **Linear SVM và Neural Net:** Hai mô hình này đạt được độ chính xác cao nhất (0.975). Điều này cho thấy khả năng học các đặc trưng phức tạp của hình ảnh của chúng.
- **Random Forest và Decision Tree:** Các mô hình dựa trên cây quyết định cũng đạt được độ chính xác khá cao (0.925 và 0.9). Chúng có ưu điểm là dễ hiểu và giải thích.
- **Naive Bayes:** Mô hình này cũng đạt được độ chính xác khá tốt (0.95), nhưng dựa trên giả định về độc lập giữa các thuộc tính, có thể không phù hợp với dữ liệu hình ảnh có nhiều tương quan.
- **Nearest Neighbors, RBF SVM và AdaBoost:** Các mô hình này có độ chính xác thấp hơn so với các mô hình còn lại.

Nhận xét về mô hình có độ chính xác cao nhất (Linear SVM và Neural Net)

- **Linear SVM:**
  - **Ưu điểm:** Đạt được độ chính xác cao, hiệu quả với dữ liệu có chiều cao.
  - **Nhược điểm:** Có thể gặp khó khăn khi dữ liệu không tuyến tính.

- **Neural Net:**
  - **Ưu điểm:** Khả năng học các đặc trưng phức tạp, đạt được độ chính xác cao trên nhiều bài toán.
  - **Nhược điểm:** Cần lượng dữ liệu lớn để huấn luyện, dễ bị overfitting nếu không điều chỉnh hyperparameter cẩn thận.

Giải thích tại sao Linear SVM và Neural Net đạt được độ chính xác cao

- **Linear SVM:** Mô hình này tìm kiếm một siêu phẳng tốt nhất để phân tách các lớp dữ liệu. Với dữ liệu hình ảnh, các đặc trưng thường có thể được biểu diễn trong không gian vector cao chiều, và Linear SVM có khả năng tìm ra các siêu phẳng phức tạp để phân loại chính xác các hình ảnh.
- **Neural Net:** Mô hình này mô phỏng cách hoạt động của não người, với nhiều lớp neuron kết nối với nhau. Các lớp đầu tiên học các đặc trưng đơn giản, trong khi các lớp sau học các đặc trưng phức tạp hơn. Nhờ cấu trúc này, Neural Net có khả năng học các đặc trưng trừu tượng và phức tạp của hình ảnh, dẫn đến độ chính xác cao.

2.3 CÁC MÔ HÌNH CHẠY TRÊN TẬP DỮ LIỆU KẾT HỢP

2.3.1 Kết quả chạy thử nghiệm

Kết quả chạy thử nghiệm với alpha combined_ranking = 0.3

Text / Img	Nearest Neighbors	Linear SVM	RBF SVM	Decision Tree	Random Forest	Neural Net	Ada Boost	Naive Bayes
Nearest Neighbors	0.775	0.675	0.675	0.7	0.65	0.65	0.7	0.675
Linear SVM	0.975	1.0	0.975	0.95	0.975	1.0	1.0	1.0
RBF SVM	0.925	0.925	0.975	0.675	0.5	0.825	0.875	0.975
Decision Tree	0.5	0.5	0.85	0.85	0.8	0.5	0.85	0.85
Random Forest	0.875	0.8	0.825	0.675	0.85	0.975	0.975	0.75
Neural Net	0.975	0.95	0.975	1.0	1.0	0.975	1.0	0.975
AdaBoost	0.9	0.9	0.5	0.9	0.5	0.5	0.8	0.85
Naive Bayes	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95

Các cặp mô hình đạt độ chính xác cao nhất (1.0)

Có 6 cặp mô hình đạt độ chính xác 1.0, bao gồm:

- **Linear SVM - Linear SVM:** Cặp này đạt được độ chính xác cao nhất. Điều này có thể giải thích bởi việc sử dụng cùng một loại mô hình cho cả dữ liệu văn bản và hình ảnh, cho phép chúng học được các đặc trưng chung và bổ sung cho nhau.
- **Linear SVM - Neural Net:** Sự kết hợp giữa SVM tuyến tính và mạng neural cho thấy hiệu quả cao. SVM có thể học được các siêu phẳng phân cách tuyến tính hiệu quả, trong khi mạng neural có khả năng học các đặc trưng phức tạp hơn.
- **Neural Net - Neural Net:** Việc sử dụng hai mạng neural để xử lý dữ liệu văn bản và hình ảnh riêng biệt, sau đó kết hợp kết quả cũng đạt được độ chính xác cao. Điều này cho thấy khả năng học sâu của mạng neural trong việc trích xuất thông tin từ cả hai loại dữ liệu.
- **Linear SVM - Neural Net:** Cặp này cũng đạt được độ chính xác cao, tương tự như cặp trên.

- **Neural Net - AdaBoost:** Sự kết hợp giữa mạng neural và AdaBoost cho thấy hiệu quả tốt. AdaBoost có thể tăng cường khả năng phân loại của mạng neural bằng cách tập trung vào các mẫu dữ liệu khó phân loại.
- **Naive Bayes - Naive Bayes:** Mặc dù đơn giản nhưng cặp này vẫn đạt được độ chính xác cao. Điều này có thể do dữ liệu được chuẩn bị tốt và các giả định của Naive Bayes phù hợp với bài toán.

Giải thích về phương pháp kết hợp và độ chính xác cao

Phương pháp kết hợp dữ liệu văn bản và hình ảnh với alpha combined_ranking = 0.3 có nghĩa là thông tin từ cả hai loại dữ liệu đều được cân nhắc, nhưng thông tin từ văn bản có thể được ưu tiên hơn một chút. Điều này có thể giải thích tại sao các cặp mô hình sử dụng Linear SVM (một mô hình thường hiệu quả với dữ liệu văn bản) lại đạt được kết quả cao.

CHƯƠNG 3 KẾT LUẬN

3.1 KẾT QUẢ ĐẠT ĐƯỢC

Các mô hình máy học đã chứng minh được khả năng phân loại hiệu quả trên cả dữ liệu văn bản và hình ảnh. Tính đến thời điểm này, phần lớn các mô hình đều cho thấy khả năng phân loại đáng kể trong một khoảng thời gian hợp lý. Từ các mô hình đơn giản như Naïve Bayes cho đến những mô hình phức tạp như Neural Net, đều đạt được hiệu suất đáng chú ý. Đặc biệt, các mô hình này được thiết kế để linh hoạt và có thể tích hợp vào nhiều hệ thống và ứng dụng khác nhau để phục vụ nhiều mục đích sử dụng khác nhau. Kết quả của nghiên cứu này đã minh chứng sự tiềm năng lớn của các mô hình máy học trong việc xử lý và phân tích các tập dữ liệu đa dạng.

3.2 HƯỚNG PHÁT TRIỂN

- Các mô hình trong nghiên cứu này có thể được tối ưu hóa thông qua điều chỉnh các tham số, sử dụng các kỹ thuật giảm overfitting hoặc tăng cường số lớp.
- Sử dụng transfer learning từ các mô hình đã được huấn luyện trước cũng là một cách hiệu quả để tiết kiệm thời gian huấn luyện.
- Mở rộng tập dữ liệu huấn luyện bằng cách đa dạng hóa các loại dữ liệu có thể giúp cải thiện hiệu suất tổng thể của mô hình.
- Việc tiếp tục áp dụng và nghiên cứu các mô hình máy học mới, đặc biệt là các kiến trúc mạng học sâu và biến thể của chúng, cũng là một hướng phát triển quan trọng.