ĐẠI HỌC BÁCH KHOA HÀ NỘI

ĐỒ ÁN TỐT NGHIỆP

Đánh giá chất lượng phán đoán của mô hình học máy khi không dùng nhãn

PHAN VĂN ĐẠT
dat.pv200130@sis.hust.edu.vn

Ngành: Khoa học máy tính

Giảng viên hướng dẫn: PGS.TS. Thân Quang Khoát _____

Chữ kí GVHD

Khoa: Khoa học máy tính

Trường: Công nghệ Thông tin và Truyền thông

HÀ NỘI, 06/2024

LỜI CẢM ƠN

Đầu tiên, em xin gửi lời cảm ơn chân thành nhất đến PGS.TS Thân Quang Khoát, em cảm ơn thầy đã luôn có những đóng góp và hướng dẫn tận tình cho em trong suốt quá trình thực hiện đồ án cũng như hoạt động ở DSLab. Em xin cảm ơn đến tất cả cán bộ giảng viên của Đại học Bách Khoa Hà Nội nói chung và trường Công nghệ thông tin và truyền thông nói riêng vì những kiến thức và kinh nghiệm mà em đã nhận được từ thầy cô.

Em cảm ơn những người anh chị, bạn bè đã đồng hành và giúp đỡ trong suốt quá trình học đại học, đặc biệt với tất cả thành viên DSLab bởi những lời khuyên tận tình. Cảm ơn em Đoàn Thế Vinh đã cộng tác và hỗ trợ rất nhiều trong quá trình thực hiện đồ án.

Cuối cùng, con xin cảm ơn bố mẹ rất nhiều vì đã luôn bên cạnh, hỗ trợ và động viên con trong suốt quá trình học tập và trưởng thành.

Cảm ơn tất cả mọi người!

LỜI CAM KẾT

Họ và tên sinh viên: Phan Văn Đạt
Điện thoại liên lạc: 0949446398
Email: dat.pv200130@sis.hust.edu.vn
Lớp: Khoa học máy tính 04
Hệ đào tạo: Cử nhân chính quy

Tôi – *Phan Văn Đạt* – cam kết Đồ án Tốt nghiệp (ĐATN) là công trình nghiên cứu của bản thân tôi dưới sự hướng dẫn của *PGS.TS Thân Quang Khoát*. Các kết quả nêu trong ĐATN là trung thực, là thành quả của riêng tôi, không sao chép theo bất kỳ công trình nào khác. Tất cả những tham khảo trong ĐATN – bao gồm hình ảnh, bảng biểu, số liệu, và các câu từ trích dẫn – đều được ghi rõ ràng và đầy đủ nguồn gốc trong danh mục tài liệu tham khảo. Tôi xin hoàn toàn chịu trách nhiệm với dù chỉ một sao chép vi phạm quy chế của nhà trường.

Hà Nội, ngày tháng năm

Tác giả ĐATN

Họ và tên sinh viên

TÓM TẮT NỘI DUNG ĐỒ ÁN

Trong thời đại công nghệ số phát triển mạnh mẽ, sự gia tăng nhanh chóng của dữ liệu đã mang lại cả cơ hội và thách thức cho lĩnh vực học máy. Học máy hiện đóng vai trò then chốt trong việc giải quyết nhiều vấn đề phức tạp như nhận dạng hình ảnh, xử lý ngôn ngữ tự nhiên, dự đoán tài chính và chăm sóc sức khỏe. Tuy nhiên, một trong những khó khăn lớn nhất là xử lý dữ liệu không nhãn, vì việc thu thập và gán nhãn dữ liệu thường tốn kém và mất nhiều thời gian. Do đó, việc tìm cách sử dụng dữ liệu không nhãn một cách hiệu quả là rất quan trọng. Vì vậy, mục tiêu của đồ án này là sẽ tập trung vào việc phát triển phương pháp đánh giá hiệu quả của các mô hình học máy khi chỉ có dữ liệu không nhãn. Đồ án đã đề xuất một độ đo mới và thực nghiệm cho thấy độ đo có khả năng đánh giá tốt mô hình khi không sử dụng nhãn. Ngoài ra, đồ án cũng đề xuất một quy trình để áp dụng các độ đo đánh giá này vào thực nghiệm một cách hiệu quả. Kết quả của đồ án sẽ đóng góp vào việc đánh giá mô hình tốt hơn khi việc gán nhãn cho dữ liệu gặp nhiều khó khăn.

Sinh viên thực hiện

(Ký và ghi rõ họ tên)

MỤC LỤC

CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI 1

1.1 Đặt vấn đề 1

1.2 Các giải pháp hiện tại và hạn chế 2

1.3 Mục tiêu và định hướng giải pháp 2

1.4 Đóng góp của đồ án 3

1.5 Bố cục đồ án 3

CHƯƠNG 2. NỀN TẢNG LÝ THUYẾT 4

2.1 Đánh giá mô hình 4

2.2 Các kết quả nghiên cứu tương tự 5

2.3 Liên tục Lipschitz 6

2.4 Một số phương pháp đánh giá mô hình bằng dữ liệu không nhãn 6

2.4.1 Meta-Distribution Energy 6

2.4.2 Average Thresholded Confidence 7

CHƯƠNG 3. PHƯƠNG PHÁP ĐỀ XUẤT 9

3.1 Tổng quan giải pháp 9

3.2 Khả năng tổng quát hóa của mô hình sử dụng dữ liệu không nhãn 9

3.3 Quy trình đánh giá mô hình bằng dữ liệu không nhãn 11

CHƯƠNG 4. PHÂN TÍCH LÝ THUYẾT 13

4.1 Chứng minh 13

4.1.1 Chứng minh biểu thức 3.1 13

4.1.2 Chứng minh biểu thức 3.2 15

4.2 Tính chất liên tục Lipschitz của ℓ_1 18

CHƯƠNG 5. ĐÁNH GIÁ THỰC NGHIỆM 19

5.1 Độ đo và phương pháp đánh giá 19

5.2 Phương pháp thí nghiệm 20
5.2.1 Bộ dữ liệu 20
5.2.2 Các phương pháp cơ sở 22
5.2.3 Kịch bản thử nghiệm 22
5.3 Kết quả thử nghiệm 25
5.3.1 Kết quả thử nghiệm trên bộ dữ liệu CIFAR-10 26
5.3.2 Kết quả thử nghiệm trên bộ dữ liệu CIFAR-10 Corrupted 26
5.3.3 Kết quả thử nghiệm trên bộ dữ liệu ImageNet1K 27
CHƯƠNG 6. KẾT LUẬN 32
6.1 Kết luận 32
6.2 Hướng phát triển trong tương lai 32
TÀI LIỆU THAM KHẢO 34

DANH MỤC HÌNH VẼ

Hình 3.1 Quy trình đánh giá mô hình 11

Hình 5.1 Các trường hợp khác nhau của độ tương quan Pearson 20
Hình 5.2 Minh họa bộ dữ liệu CIFAR-10 21
Hình 5.3 Minh họa các biến đổi và nhiễu trong CIFAR-10-C 22
Hình 5.4 Minh họa bộ dữ liệu ImageNet1K 23

DANH MỤC BẢNG BIỂU

Bảng 5.1 Số lượng ảnh trong mỗi tập của bộ dữ liệu ImageNet1K. . . . 21
Bảng 5.2 Độ tương quan Pearson đối với độ chính xác trên tập không nhãn CIFAR-10 của các độ đo đánh giá. 26
Bảng 5.3 Độ tương quan Pearson đối với độ chính xác trên tập không nhãn CIFAR-10 của độ đo prob measure với các chiến lược phân cụm khác nhau. 27
Bảng 5.4 Độ tương quan Pearson đối với độ chính xác của các độ đo khác nhau trên các chỉnh sửa của CIFAR-10-C. Độ đo đề xuất được tính với chiến lược phân cụm láng giềng gần nhất với 1000 tâm. **Các mô hình được huấn luyện không dùng tăng cường dữ liệu.** . 28
Bảng 5.5 Độ tương quan Pearson đối với độ chính xác của các độ đo khác nhau trên các chỉnh sửa của CIFAR-10-C. Độ đo đề xuất được tính với chiến lược phân cụm láng giềng gần nhất với 1000 tâm. **Các mô hình được huấn luyện dùng tăng cường dữ liệu.** 29
Bảng 5.6 Các độ đo trên bộ dữ liệu ImageNet1K 30

DANH MỤC THUẬT NGỮ VÀ TỪ VIẾT TẮT

Thuật ngữ	**Ý nghĩa**
ATC	Average Thresholded Confidence
MDE	Meta-Distribution Energy

CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI

1.1 Đặt vấn đề

Trong bối cảnh hiện đại, sự phát triển vượt bậc của các công nghệ số và sự gia tăng nhanh chóng của dữ liệu đã mở ra nhiều cơ hội và thách thức trong lĩnh vực học máy. Công nghệ số không chỉ thay đổi cách chúng ta sống và làm việc, mà còn mang lại những bước tiến lớn trong việc thu thập và xử lý dữ liệu. Sự bùng nổ của Internet, mạng xã hội, và các thiết bị kết nối đã tạo ra lượng dữ liệu khổng lồ chưa từng có. Những dữ liệu này có thể đến từ nhiều nguồn khác nhau như văn bản, hình ảnh, video, và các tín hiệu cảm biến, tạo nên một nguồn tài nguyên vô cùng phong phú và đa dạng cho các ứng dụng học máy. Từ đó, học máy đã trở thành một công cụ quan trọng trong việc giải quyết nhiều bài toán phức tạp từ nhận dạng hình ảnh, xử lý ngôn ngữ tự nhiên cho đến dự đoán tài chính và chăm sóc sức khỏe. Đặc biệt, với sự phát triển về khả năng tính toán của phần cứng, học sâu (deep learning) cũng được quan tâm nghiên cứu và cho ra những kết quả vượt trội ở các lĩnh vực. Các mô hình học sâu, với khả năng tự động trích xuất đặc trưng từ dữ liệu thô, đã đạt được nhiều thành tựu ấn tượng trong các bài toán phức tạp như nhận dạng hình ảnh, xử lý ngôn ngữ tự nhiên hay xử lý giọng nói. Tuy nhiên, việc đào tạo các mô hình học sâu thường đòi hỏi một lượng lớn dữ liệu và tài nguyên tính toán, điều này đặt ra một trong những thách thức lớn nhất của học máy cũng như học sâu là vấn đề dữ liệu không nhãn (unlabeled data).

Dữ liệu không nhãn là loại dữ liệu mà các đầu vào không có các nhãn hay thông tin đầu ra tương ứng. Đây là loại dữ liệu rất phổ biến trong thực tế, khi mà các thông tin được thu thập một cách tự nhiên từ nhiều nguồn khác nhau như văn bản, hình ảnh, âm thanh, hay dữ liệu cảm biến, nhưng không có nhãn hoặc chú thích đi kèm. Trong khi dữ liệu có nhãn (labeled data) giúp mô hình học máy học một cách rõ ràng từ các ví dụ, đánh giá các mô hình học máy thông qua dữ liệu có nhãn cũng không gặp nhiều trở ngại, thì việc thu thập và gán nhãn cho dữ liệu thường rất tốn kém và đòi hỏi nhiều công sức. Để tạo ra một tập dữ liệu có nhãn lớn và chất lượng, cần phải có sự can thiệp của con người để gán nhãn chính xác cho từng mục dữ liệu. Quá trình này không chỉ đòi hỏi thời gian và chi phí mà còn yêu cầu sự hiểu biết sâu sắc về lĩnh vực ứng dụng để đảm bảo các nhãn được gán một cách chính xác và nhất quán. Do đó, để giảm thiểu chi phí cho gán nhãn dữ liệu, việc sử dụng hiệu quả dữ liệu không nhãn trở thành một vấn đề quan trọng cần được giải quyết.

Trong đồ án tốt nghiệp này, em sẽ tập trung vào việc phát triển phương pháp để đánh giá các mô hình học máy khi chỉ có sẵn dữ liệu không nhãn, đánh giá và so