TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

ĐỀ CƯƠNG KHOÁ LUẬN TỐT NGHIỆP

HỌC CÓ GIÁM SÁT VỚI DỮ LIỆU CÓ PHÂN BỐ THAY ĐỔI BẰNG MÔ HÌNH DỰA TRÊN QUAN HỆ NHÂN QUẢ

(Supervised learning with data distribution shift using causality-based model)

1 THÔNG TIN CHUNG

Người hướng dẫn:

– Thầy Trần Trung Kiên

– Cô Nguyễn Ngọc Thảo (Khoa Công nghệ Thông tin)

Nhóm sinh viên thực hiện:

1. Phan Trường An (MSSV: 20120032)
2. Phạm Dương Trường Đức (MSSV: 20120061)

Loại đề tài: Nghiên cứu

Thời gian thực hiện: Từ *01/2024* đến *06/2024*

2 NỘI DUNG THỰC HIỆN

2.1 Giới thiệu về đề tài

Học có giám sát là một loại của học máy mà máy tính sẽ được huấn luyện trên một bộ dữ liệu được gán nhãn (có nghĩa là dữ liệu đầu vào sẽ tương ứng với dữ liệu đầu ra mong muốn). Máy tính sẽ học trên bộ dữ liệu đó và đưa ra dự đoán cho dữ liệu đầu vào mới. Ví dụ, để dự đoán giá của một ngôi nhà mới, một mô hình học máy sẽ được huấn luyện trên bộ dữ liệu bao gồm thông tin (vị trí, diện tích, số phòng, số tầng,...) của các ngôi nhà được bán trước đó đại diện cho đầu vào và giá của chúng đại diện cho đầu ra (nhãn). Sau quá trình huấn luyện, khi đưa thông tin của ngôi nhà mới vào mô hình, mô hình sẽ dự đoán được giá của ngôi nhà đó.

Dữ liệu có phân bố thay đổi (Data distribution shift) là một hiện tượng trong học có giám sát xảy ra khi dữ liệu mà mô hình phải dự đoán có sự thay đổi phân bố so với dữ liệu mà mô hình được huấn luyện; điều này khiến cho độ chính xác của kết quả dự đoán của mô hình giảm.

Gọi X là không gian dữ liệu đầu vào và Y là không gian dữ liệu đầu ra (nhãn) của một mô hình. Trong học có giám sát, dữ liệu huấn luyện của mô hình có thể được xem như một tập hợp các mẫu từ phân bố kết hợp $P(X, Y)$. Phân bố này có thể được biến đổi như sau:

$$P(X, Y) = P(X)P(Y|X)$$

Từ phép biến đổi trên, ta có hai loại thay đổi phân bố:

- **Covariate shift** là khi $P(X)$ thay đổi, nhưng $P(Y|X)$ không thay đổi.
- **Concept shift** là khi $P(Y|X)$ thay đổi, nhưng $P(X)$ không thay đổi.

Tại đề tài khóa luận này chúng em tập trung vào vấn đề **Thay đổi miền (Domain shift)** là một nhóm con thuộc **Covariate shift**.

Bài toán "học có giám sát với dữ liệu có phân bố thay đổi" mà chúng em sẽ làm trong khóa luận được phát biểu như sau:

- Cho dữ liệu của K $(K \geq 1)$ miền huấn luyện $D_{Train} = \{D^i | i = 1, ..., K\}$ với $D^i = (x_j^i, y_j^i)_{j=1}^{n_i}, x \in X, y \in Y$ là thể hiện dữ liệu của miền thứ i. Phân bố đầu vào của các miền là khác nhau, nghĩa là $P_X^i \neq P_X^j, 1 \leq i \neq j \leq K$.
- Mục tiêu của bài toán là xây dựng một hàm dự đoán $h : X \longrightarrow Y$ từ dữ liệu của K miền huấn luyện được cho để độ lỗi khi dự đoán trên dữ liệu của miền mục tiêu D_{Test} là nhỏ nhất. Nghĩa là hàm dự đoán có thể khái quát và chống chịu tốt với sự thay đổi miền. D_{Test} là dữ liệu thỏa điều kiện $P_X^{Test} \neq P_X^i, i \in \{1, ..., K\}$.

Dữ liệu thực tế thường có xuất hiện cùng lúc nhiều loại phân bố thay đổi và phân bố thay đổi của dữ liệu có thể xảy ra trên nhiều thuộc tính. Đây là một thách thức cho các mô hình học máy. Bởi vì một mô hình dự đoán có độ chính xác cao với loại phân bố thay đổi này sẽ có độ chính xác thấp với loại phân bố thay đổi khác. Hơn nữa, các mô hình giải quyết bài toán "học có giám sát với dữ liệu có phân bố thay đổi" thường chỉ tập trung vào dữ liệu thay đổi trên một thuộc tính. Những mô hình này sẽ có độ chính xác thấp khi gặp dữ liệu có phân bố thay đổi trên nhiều thuộc tính.

Nếu giải quyết được bài toán này, ta có thể giúp các mô hình học máy khái quát tốt được các miền, từ đó giúp mô hình chống chịu tốt với sự thay đổi phân bố dữ liệu, nhờ vậy các mô hình sẽ không bị giảm hiệu suất và đưa ra dự đoán đủ tốt theo yêu cầu của người dùng. Việc chống chịu tốt với sự thay đổi miền còn giúp mở rộng ứng dụng của các mô hình học máy trong thực tế.

Qua quá trình tìm hiểu, chúng em nhận thấy hướng tiếp cận sử dụng mô hình dựa trên quan hệ nhân quả là một hướng tiếp cận tiềm năng. Vì vậy, chúng em sẽ tập trung tìm hiểu sâu hướng tiếp cận này trong khóa luận.

2.2 Mục tiêu đề tài

- Tìm hiểu bài toán "học có giám sát với dữ liệu có phân bố thay đổi" (tình hình nghiên cứu, các phương pháp tiếp cận/giải quyết). Từ đó, chọn ra một phương pháp tiềm năng để hiểu sâu.
- Cài đặt thành công phương pháp được đề xuất trong bài báo mà chúng em tham khảo và đạt được kết quả như bài báo đó. Tiến hành cài đặt các phương pháp khác để so sánh.
- Tìm hiểu, thí nghiệm để hiểu thêm về ưu/nhược điểm của phương pháp đã chọn. Sau đó, xem xét và cải tiến nếu có thể.
- Rèn luyện kỹ năng nghiên cứu, mở rộng kiến thức liên quan đến bài toán.
- Nâng cao khả năng tự học, làm việc nhóm, kỹ năng trình bày,...

2.3 Phạm vi của đề tài

Trong đề tài này, chúng em sẽ tìm hiểu và cài đặt lại phương pháp từ một bài báo uy tín. Trong quá trình cài đặt, nếu có thời gian và khả năng, chúng em sẽ tiến hành cải tiến phương pháp. Theo chúng em thì việc hiểu rõ cơ sở lý thuyết và cài đặt được phương pháp sẽ cần nhiều thời gian, nên chúng em xem đây là mục tiêu chính cần tập trung thực hiện. Khi đã hiểu rõ và cài đặt thành công phương pháp của bài báo, chúng em sẽ tiến hành các thử nghiệm để tìm ra ưu/khuyết điểm của phương pháp đó.
Trong đề tài này, chúng em sẽ thử nghiệm phương pháp với bài toán phân loại. Đề tài chủ yếu làm về dữ liệu ảnh. Chúng em dự định sẽ thực hiện với 3 bộ dữ liệu là MNIST, small NORB và Waterbirds.

2.4 Cách tiếp cận dự kiến

Dưới đây là một số phương pháp giải quyết bài toán mà chúng em đã tìm hiểu cho đến thời điểm hiện tại.

- Một phương pháp thường được sử dụng là Correlation Alignment (CORAL) được đề xuất bởi B. Sun và cộng sự trong bài báo "Return of Frustratingly Easy Domain Adaptation" [1]. Phương pháp này điều chỉnh hiệp phương sai của dữ liệu từ miền huấn luyện D_{Train} theo dữ liệu của miền mục tiêu D_{Test}. Từ đó, dữ liệu từ D_{Train} sẽ có phân bố tương đồng với dữ liệu từ D_{Test}. Nhờ vậy mà mô hình huấn luyện được sẽ khái quát hóa tốt hơn. Phương pháp này đòi hỏi phải có được dữ liệu từ D_{Test}.

- Các phương pháp tăng cường dữ liệu là một trong những hướng được sử dụng phổ biến trong việc khái quát miền. Với ý tưởng là dữ liệu càng lớn thì mô hình sẽ càng có nhiều sự đa dạng dữ liệu để khái quát các miền. Mixup [2] do Zhang và cộng sự đề xuất là một phương pháp sinh dữ liệu thường được sử dụng để tăng cường dữ liệu trong bài toán khái quát miền. Mixup sinh dữ liệu mới bằng cách thực hiện nội suy tuyến tính giữa hai đầu vào bất kỳ và nhãn tương ứng của chúng trong bộ dữ liệu với một trọng số được lấy mẫu từ phân bố Beta. Đối với phương pháp Mixup nói riêng và các phương pháp tăng cường dữ liệu nói chung, đôi khi sự tăng cường dữ liệu sinh ra những đặc trưng không cần thiết trong việc khái quát miền, khiến cho tốn tài nguyên mà không giải quyết được yêu cầu bài toán.

- Một số các chiến lược học cũng có thể giải quyết được yêu cầu khái quát miền như là Esemble learning, Meta-learning, Gradient operation,... [3] Trong đó meta-learning là một trong các chiến lược được sử dụng và nghiên cứu phổ biến. MLDG [4] (meta-learning for domain generaliztion) sử dụng chiến lược này trong việc khái quát miền. Ý tưởng của MLDG là chia dữ liệu trong các miền huấn luyện D_{Train} thành các bộ dữ liệu meta-train và meta-test để mô phỏng sự thay đổi miền, từ đó học được biểu diễn của các miền và khái quát chúng. Điểm yếu của phương pháp này là nếu dữ liệu trong các miền huấn luyện không đa dạng thì khó có thể khái quát tốt được trên miền mục tiêu và

có thể dẫn đến hiện tượng overfitting trên các đặc điểm cụ thể của miền huấn luyện.

- Trong bài toán khái quát miền, các phương pháp dựa trên mối quan hệ nhân quả là hướng tiếp cận tiềm năng. Các mô hình học máy thông thường có thể dựa vào các biến gây nhiễu trong dữ liệu huấn luyện để đưa ra quyết định khiến cho dự đoán không chính xác với dữ liệu mới của miền mục tiêu (ví dụ trong bài toán nhận diện con vật, các hình ảnh về bò trong dữ liệu huấn luyện thường xuất hiện cùng với nền cỏ; nên nếu ta đưa một hình ảnh con bò ở bãi biển thì mô hình có thể dự đoán không chính xác). Các mô hình học dựa trên mối quan hệ nhân quả có thể chỉ ra những nguyên nhân thật sự dẫn đến nhãn đúng, chứ không bị ảnh hưởng bởi các biến gây nhiễu trong dữ liệu huấn luyện. Nhờ đó, các mô hình này có thể dự đoán tốt với dữ liệu mới của miền mục tiêu. Năm 2022, Kaur và cộng sự đã công bố bài báo "Modeling the data-generating process is necessary for out-of-distribution generalization" ở hội nghị International Conference on Learning Representations [5]. Mô hình được đề xuất đã rút trích ra các ràng buộc chính xác từ đồ thị nhân quả để tìm ra tất cả các nguyên nhân dẫn đến nhãn Y. Từ đó, mô hình có thể khái quát tốt được các miền và đưa ra dự đoán chính xác nhất. Ngoài ra, mô hình của Kaur và cộng sự [5] còn có thể khái quát tốt được sự thay đổi đa thuộc tính trên các miền.

Trong khóa luận này, chúng em quyết định tập trung tìm hiểu và cài đặt lại phương pháp theo thuật toán được đề xuất trong bài báo [5]. Phương pháp này có độ chính xác cao hơn các phương pháp còn lại trong nhiều trường hợp. Hơn nữa, nó còn có thể giải quyết bài toán dữ liệu có phân bố thay đổi trên nhiều thuộc tính. Trong khi các phương pháp khác chỉ tập trung vào dữ liệu có phân bố thay đổi trên một thuộc tính. Ngoài ra, học dựa trên mối quan hệ nhân quả có thể được ứng dụng mở rộng ra rất nhiều bài toán. Tuy nhiên, kiến thức nền tảng của học dựa trên

mối quan hệ nhân quả theo chúng em nghĩ là không dễ để có thể làm chủ. Chúng em xác định hiểu rõ bài báo này cũng sẽ là cơ hội để hiểu rõ các kiến thức học dựa trên mối quan hệ nhân quả.

2.5 Kết quả dự kiến của đề tài

Qua khóa luận này, chúng em mong muốn đạt được các kết quả sau:

- Cài đặt từ đầu phương pháp được đề xuất và đạt được các kết quả đầu ra như trong bài báo [5].
- Có các thử nghiệm để hiểu rõ tiềm năng phát triển của hướng tiếp cận này cũng như hạn chế của nó.
- Thử nghiệm phương pháp trên các bài toán hồi quy, hoặc trên các bộ dữ liệu văn bản và cải tiến phương pháp nếu có đủ thời gian.

2.6 Kế hoạch thực hiện

Thời gian	Công việc	Người thực hiện
Từ 01/01/2024 đến 31/01/2024	- Tìm hiểu các bài báo khái quát miền dựa trên học nhân quả và chọn ra một phương pháp tiềm năng. - Lấy dữ liệu được sử dụng trong bài báo đó.	Phan Trường An Phạm Dương Trường Đức

Từ 01/02/2024 đến 15/03/2024	- Tìm hiểu một số bài báo giải quyết vấn đề khái quát miền và chọn ra một số phương pháp để so sánh với bài báo chính. - Viết đề cương thực hiện khóa luận. - Nắm rõ ý tưởng thực hiện của bài báo chính.	Phan Trường An Phạm Dương Trường Đức
Từ 16/03/2024 đến 15/05/2024	- Hiểu sâu về kiến thức nền tảng của phương pháp học dựa trên quan hệ nhân quả. - Hiểu sâu bài báo chính. - Cài đặt lại phương pháp và tiến hành các thí nghiệm để tái hiện kết quả trong bài báo chính. - Tiến hành các thí nghiệm mở rộng (nếu còn thời gian).	Phan Trường An Phạm Dương Trường Đức
Từ 16/05/2024 đến 31/06/2024	- Viết báo cáo khóa luận tốt nghiệp và chuẩn bị bài thuyết trình.	Phan Trường An Phạm Dương Trường Đức

Tài liệu

[1] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proceedings of the AAAI conference on artificial intelligence*, 2016.

[2] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *ICLR*, 2017.

[3] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. Yu, "Generalizing to unseen domains: A survey on domain generalization," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[4] D. Li, Y. Yang, Y.-Z. Song, and T. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *AAAI*, 2018.

[5] J. N. Kaur, E. Kiciman, and A. Sharma, "Modeling the data-generating process is necessary for out-of-distribution generalization," *International Conference on Learning Representations*, 2022.

TP. Hồ Chí Minh, ngày 04 tháng 04 năm 2024

XÁC NHẬN
CỦA NGƯỜI HƯỚNG DẪN
(Ký và ghi rõ họ tên)

Trần Trung Kiên

NHÓM SINH VIÊN THỰC HIỆN
(Ký và ghi rõ họ tên)

Phan Trường An

Nguyễn Ngọc Thảo

Phạm Dương Trường Đức