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

ĐỀ CƯƠNG KHOÁ LUẬN TỐT NGHIỆP

KẾT HỢP CÁC ĐỐI TƯỢNG VÀ ĐẶC TRƯNG LÂN CẬN CHO BÀI TOÁN PHÂN LỚP ĐA NHÃN

Combining instance and feature neighbours for multi-label classification

1 THÔNG TIN CHUNG

Người hướng dẫn:

– GS.TS Lê Hoài Bắc (Khoa Công nghệ Thông tin)

Nhóm Sinh viên thực hiện:

1. Trần Xuân Quý (MSSV: 18120231)
2. Trần Hữu Chí Bảo (MSSV: 18120288)

Loại đề tài: Nghiên cứu

Thời gian thực hiện: Từ *09/2021* đến *03/2022*

2 NỘI DUNG THỰC HIỆN

2.1 Giới thiệu về đề tài

Bắt nguồn từ yêu cầu giải quyết bài toán phân loại văn bản - với mỗi văn bản thường thuộc nhiều hơn một phân lớp - và bài toán chẩn đoán y khoa, gần đây

ngày càng nhiều ứng dụng, bài toán có nhu cầu sử dụng các thuật toán phân lớp đa nhãn, như là phân loại chức năng protein, thể loại âm nhạc và ngữ nghĩa khung cảnh của ảnh [1]. Hai hướng tiếp cận chính giải quyết các bài toán phân đa lớp hiện có gồm chia nhỏ bài toán thành tổ hợp những tác vụ phân lớp đơn nhãn và chỉnh sửa các thuật toán phân lớp đơn sao cho phù hợp với tác vụ đa lớp. Một số mô hình gần đây xét thêm yếu tố phụ thuộc tiềm tàng cũng như là phân bố lệch giữa các nhãn sử dụng phương pháp giảm chiều không gian hay xây dựng nhóm các mô hình cây có thứ bậc. Cách tiếp cận này cho độ chính xác và thời gian thực thi nhanh khi kiểm tra nhưng đòi hỏi lượng tài nguyên cho phần huấn luyện là rất lớn, cũng như yêu cầu phải tối ưu nhiều siêu tham số [2, 3].

Ở đề tài này nhóm thực hiện tìm hiểu và cài đặt các thuật toán k lân cận gần nhất có áp dụng một số kết quả ở lĩnh vực hệ thống tư vấn, cụ thể là thuật toán k lân cận gần nhất dựa trên đối tượng và đặc trưng, lần lượt áp dụng các tính năng của thuật toán lọc cộng tác dựa trên người dùng và dựa trên từng mục, và sau cùng là tổ hợp tuyến tính của 2 thuật toán trên. Nhìn chung 2 thuật toán k lân cận trên được đánh giá là hiệu quả hơn về mặt tính toán so với những mô hình xuất hiện trước cũng như có thể hoạt động tốt với phần cứng thông dụng.

2.2 Mục tiêu đề tài

Đề tài tập trung vào thực hiện việc tìm mối quan hệ sự phụ thuộc có thể có giữa các nhãn, đối phó với độ lệch nhãn bởi vì nhãn chỉ tập trung ở một số đối tượng. Ngoài ra, cải thiện chi phí, tốc độ cũng như tăng độ chính xác để tạo ra 1 mô hình khi số lượng nhãn rất lớn.

Cụ thể sử dụng thuật toán dựa vào k lân cận dựa trên đối tượng gần nhất tiếp sau đó là thuật toán k lân cận gần nhất dựa trên đặc trưng. Và cuối cùng sử dụng thuật toán sinh cắt tỉa [2] bằng cách kết hợp cả 2 phương pháp dựa trên đối tượng và đặc trưng.

Việc áp dụng ba thuật toán gặp phải một số thách thức như kích thước lớn, nhiều thuộc tính đặc trưng và nhãn gây khó khăn cho việc xử lý dữ liệu, phân tích và kiểm tra lại kết quả. Từ việc đánh giá hiệu quả cũng như mức độ ảnh hưởng đến

việc xây dựng mô hình, nhằm tìm ra ưu điểm và nhược điểm của từng thuật toán. Từ đó, biết cách kết hợp chúng sao cho đạt hiệu quả cao nhất mức độ đa dạng của các đối tượng để chọn một mô hình phù hợp nhất.

2.3 Phạm vi của đề tài

Trong phạm vi đề tài này, đầu tiên nhóm tập trung vào việc cài đặt và tái hiện lại kết quả - bao gồm kết quả các độ đo đánh giá cũng như thời gian chạy chương trình - của 2 thuật toán k lân cận gần nhất và tổ hợp tuyến tính giữa chúng cùng với các độ đo đánh giá liên quan. Để kiểm tra việc tái hiện, nhóm cho chạy cài đặt của mình trên các tập dữ liệu đã được sử dụng ở bài báo gốc như là Medical, Delicious, Eurlex, AmazonCat, ... [2, 4, 5, 6]. Khi đã hoàn thành yêu cầu trên, các thành viên tiếp tục tìm hiểu những phương pháp có thể áp dụng giúp cho mô hình dự đoán có thể đạt được kết quả tốt hơn theo những độ đo đã được nêu.
Những tập dữ liệu được sử dụng trong đề tài này đều là những tập dữ liệu văn bản, tuy nhiên việc tiền xử lý những đoạn văn bản sang định dạng có thể áp dụng vào mô hình dự đoán không nằm trong phạm vi đề tài. Do đó các tập dữ liệu được sử dụng đều ít nhất đã qua tiền xử lý như là áp dụng mô hình túi từ để rút trích đặc trưng.

2.4 Cách tiếp cận dự kiến

Nhóm dự kiến trải qua các giai đoạn sau trong quá trình thực hiện đề tài:

- Giai đoạn 1: Tìm hiểu về bài toán phân lớp đa nhãn. Ở phần này nhóm thực hiện tìm hiểu về đặc điểm của tập dữ liệu lớn và cực lớn cũng như cách lưu trữ thường dùng. Thêm vào đó nhóm cũng tìm hiểu về các mô hình hiện có để giải quyết bài toán này, các hướng tiếp cận được sử dụng cũng như những gì mô hình đã đạt được và còn hạn chế. Ngoài ra nhóm cũng tìm hiểu và tóm tắt những thông tin quan trọng của mô hình mới được đề xuất.
- Giai đoạn 2: Nhóm tập trung tìm hiểu chi tiết và cài đặt các thuật toán k lân cận gần nhất dựa trên đối tượng và đặc trưng. Các thành viên tìm hiểu sâu hơn về các bước thực hiện của thuật toán trong bài báo [2] và mã nguồn của

mô hình [7], cũng như có thể thực hiện một số thay đổi (nếu có) cho phù hợp với ngôn ngữ lập trình Python.

- Giai đoạn 3: Nhóm thực hiện kết hợp 2 thuật toán k lân cận trong giai đoạn 2 và cài đặt thuật toán ngưỡng điểm dự đoán cũng như các độ đo đánh giá. Thuật toán ngưỡng điểm dự đoán thực hiện loại bỏ các nhãn có điểm dự đoán thấp hơn hay bằng ngưỡng được chọn. Việc tìm ra ngưỡng phù hợp nhất dựa trên việc tối thiểu khác biệt giữa số lượng nhãn trung bình trong tập chân trị và tập dự đoán [2]. Sau đó nhóm kiểm tra kết quả trên các độ đo đánh giá được sử dụng bởi nhóm tác giá và thời gian thực thi chương trình.
- Giai đoạn 4: Tìm hiểu và cài đặt các hướng tiếp cận có thể có để cải thiện kết quả trên các phép đo của mô hình và/hoặc thời gian huấn luyện trên các tập dữ liệu lớn và cực lớn. Cuối cùng là nhìn lại quá trình làm khóa luận, rút ra những ưu, nhược điểm của mô hình và hoàn thành cuốn luận văn cũng như bản trình bày.

2.5 Kết quả dự kiến của đề tài

Sau quá trình nghiên cứu, chúng tôi kỳ vọng đạt được các kết quả chính như sau:

- Nắm được ý tưởng xây dựng mô hình KNN dựa trên đối tượng và đặc trưng cũng như thuật toán LCIF.
- Hoàn thành việc tái hiện kết quả của mô hình ở tiêu chí kết quả trên độ đo đánh giá cũng như thời gian huấn luyện mô hình.
- Cải thiện độ chính xác và tài nguyên tính toán của cài đặt này so với bài báo gốc.
- Đề xuất giải pháp tối ưu theo hướng cải thiện hiệu suất của mô hình thông qua việc tích hợp các thuật toán dựa trên đối tượng và đặc trưng.

2.6 Kế hoạch thực hiện

Kế hoạch thực hiện đề tài cũng như phân công công việc cho mỗi thành viên như sau:

Giai đoạn	Công việc	Thời gian dự kiến	Người thực hiện
1	Tìm hiểu về bài toán, mục đích giải, mô hình được đề xuất và các tập dữ liệu	01/09/2021 - 15/09/2021	Cả hai
2	Cài đặt các thuật toán instance-based và features-based KNN sử dụng kiểu dữ liệu Python list, dictionary	15/09/2021 - 15/10/2021	Trần Hữu Chí Bảo
	Cài đặt các thuật toán instance-based và features-based KNN sử dụng thư viện scipy.sparse	15/09/2021 - 15/10/2021	Trần Xuân Quý
3	Cài đặt thuật toán LCIF và instance-knn-fast sử dụng Python list, dictionary	15/10/2021 - 30/10/2021	Trần Hữu Chí Bảo
	Cài đặt thuật toán LCIF và instance-knn-fast sử dụng scipy.spare; Cài đặt hàm chuẩn hóa dữ liệu theo dòng và cột	15/10/2021 - 30/10/2021	Trần Xuân Quý
	Kiểm tra cài đặt trên các tập dữ liệu với các độ đo đánh giá và thời gian thực thi	30/10/2021 - 15/11/2021	Cả hai
4	Tìm hiểu và cài đặt các hướng tiếp cận có thể có để cải thiện kết quả	15/11/2021 - 01/2022	Cả hai
	Nhận ý kiến từ GVHD và hoàn thiện những nội dung còn lại của khóa luận	01/2022 - 02/2022	Cả hai
	Viết cuốn luận văn, làm slide và luyện tập thuyết trình	02/2022 - 03/2022	Cả hai

Bảng 1: Bảng phân chia công việc

Tài liệu

[1] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 3, no. 3, pp. 1–13, 2007.

[2] L. Feremans, B. Cule, C. Vens, and B. Goethals, "Combining instance and feature neighbours for extreme multi-label classification," *International Journal of Data Science and Analytics*, vol. 10, no. 3, pp. 215–231, 2020.

[3] E. Gibaja and S. Ventura, "Multi-label learning: a review of the state of the art and ongoing research," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 4, no. 6, pp. 411–444, 2014.

[4] G. Tsoumakas, E. Spyromitros-Xioufis, J. Vilcek, and I. Vlahavas, "Mulan: A java library for multi-label learning," *The Journal of Machine Learning Re-*

search, vol. 12, pp. 2411–2414, 2011.

[5] J. Read, P. Reutemann, B. Pfahringer, and G. Holmes, "Meka: a multi-label/multi-target extension to weka," 2016.

[6] K. Bhatia, K. Dahiya, H. Jain, P. Kar, A. Mittal, Y. Prabhu, and M. Varma, "The extreme classification repository: Multi-label datasets and code," 2016.

[7] L. Feremans, B. Cule, C. Vens, and B. Goethals, "Lcif: Combining instance and feature neighbours for extreme multi-label classification," 2020.

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

TP. Hồ Chí Minh, 10/11/2021
NHÓM SINH VIÊN THỰC HIỆN
(Ký và ghi rõ họ tên)