

ĐÓNG GÓP CHÍNH CỦA LUẬN ÁN TIẾN SĨ KỸ THUẬT

Tên đề tài: Khai phá quan điểm với các phương pháp quần thể

Ngành: Khoa học máy tính

Mã số: 9480101

Họ và tên NCS: Trần Quang Phúc

Người hướng dẫn khoa học:

1. GS.TS. Huỳnh Xuân Hiệp

2. TS. Lê Thị Mỹ Hạnh

Cơ sở đào tạo: Trường Đại học Bách khoa – Đại học Đà Nẵng

NHỮNG ĐÓNG GÓP MỚI CỦA LUẬN ÁN

- Đề xuất mô hình Rừng phân loại quan điểm (Opinion Classification Forest – F_{CF}) dựa trên biểu diễn ma trận bộ năm quan điểm kích thước $N \times 5$ (thực thể, khía cạnh, tình cảm, người giữ quan điểm, thời gian) kết hợp entropy Shannon có giám sát làm tiêu chí phân chia nút. Cấu trúc ngữ nghĩa của bộ năm quan điểm được bảo toàn trong suốt quá trình xây dựng cây phân loại; mô hình tích hợp cơ chế giải thích qua tầm quan trọng đặc trưng (Feature Importance) và biểu đồ phụ thuộc từng phần (PDP). Theo số liệu đã công bố (chia ngẫu nhiên theo dòng bộ năm), mô hình đạt độ chính xác 0,77 trên hai tập dữ liệu yelpHotelData và yelpResData.
- Đề xuất mô hình Rừng mật độ quan điểm (Opinion Density Forest – F_{DF}) tích hợp thuật toán phân cụm DBSCAN tại các nút lá và entropy Gaussian làm tiêu chí phân chia nhằm xử lý dữ liệu mất cân bằng theo mật độ phân bố của các cụm quan điểm, đồng thời sử dụng phương pháp giải thích dựa trên giá trị Shapley (SHAP) và chỉ số tương tác Shapley để phân tích tầm quan trọng và tương tác giữa các thành phần bộ năm. Theo số liệu đã công bố, mô hình đạt độ chính xác trung bình 0,78 mà không cần GPU; SHAP xác định thực thể (Entity) là đặc trưng chủ đạo (trung bình $|\text{SHAP}| \approx 0,087$).
- Đề xuất mô hình Rừng đa tập quan điểm (Opinion Manifold Forest – F_{MF}) học trực tiếp cấu trúc đa tập thấp chiều từ tính đồng xuất hiện tại nút lá của rừng ngẫu nhiên không giám sát, thông qua ma trận quan hệ $W \in \mathbb{R}^{N \times N}$ và khoảng cách đa tập $dist_{\mathcal{M}}$, không cần giải bài toán trị riêng toàn cục như ISOMAP hay UMAP; dữ liệu được phân cụm DBSCAN trên không gian đa tập và kết quả được tổng hợp qua bộ học siêu phân tầng (Stacking meta-learner). Theo số liệu đã công bố, mô hình đạt độ chính xác trung bình cao nhất trong ba mô hình đề xuất (0,80) và diễn giải được qua ma trận quan hệ W ; chi phí lưu trữ ma trận W là rào cản khi mở rộng quy mô dữ liệu.
- Xây dựng quy trình chuyển đổi bốn bộ dữ liệu đánh giá sang cấu trúc bộ năm quan điểm chuẩn hóa (e, a, s, h, t) phục vụ huấn luyện và đánh giá các mô hình rừng đề xuất, bao gồm hai bộ dữ liệu tiếng Anh yelpHotelData và yelpResData (dữ liệu công bố của

Mukherjee và cộng sự) và hai bộ dữ liệu tiếng Việt VLSP-2018 Restaurant và VLSP-2018 Hotel (dữ liệu của cộng đồng VLSP-2018); mô hình trích xuất bộ năm quan điểm từ văn bản đạt $F1 = 0,78$ trên tập Twitter US Airline Sentiment.

- Đánh giá lại ba mô hình theo quy trình chia theo đánh giá gốc ($5 \text{ hạt giống} \times 10.000$ đánh giá): Balanced Accuracy của các mô hình chỉ dùng bộ năm đạt khoảng $0,51-0,57$, trong đó F_{CF} và F_{DF} tương đương Random Forest, F_{MF} thấp hơn Random Forest khoảng $0,02-0,03$; qua đó định lượng mức rò rỉ dữ liệu khi chia theo dòng bộ năm (Accuracy $0,946$ so với $0,611$ với cùng Random Forest) và khuyến nghị chia theo đánh giá gốc khi đánh giá mô hình trên dữ liệu bộ năm.

Đà Nẵng, ngày tháng năm 2026

Xác nhận của người hướng dẫn khoa học

Người hướng dẫn 1

Người hướng dẫn 2

Nghiên cứu sinh

GS.TS. Huỳnh Xuân Hiệp

TS. Lê Thị Mỹ Hạnh

Trần Quang Phúc