

ĐÓNG GÓP CHÍNH CỦA LUẬN ÁN TIẾN SĨ KỸ THUẬT

Tên đề tài: Dự đoán lỗi phần mềm dựa trên mã nguồn.

Chuyên ngành: Khoa học máy tính

Mã số: 9.48.01.01

Họ và tên NCS: Hà Thị Minh Phương

Người hướng dẫn khoa học:

1. PGS.TS Nguyễn Thanh Bình

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

- Luận án nghiên cứu và đề xuất các phương pháp lựa chọn đặc trưng nhằm loại bỏ những đặc trưng dư thừa hoặc không liên quan trong tập dữ liệu dự đoán lỗi phần mềm, giúp giảm nhiễu và độ phức tạp của mô hình, đồng thời nâng cao chất lượng của dữ liệu đầu vào và độ chính xác của quá trình dự đoán lỗi. Bên cạnh đó, luận án phân tích và đánh giá hiệu quả của các kỹ thuật lựa chọn đặc trưng, đồng thời đề xuất tích hợp các phương pháp này với kỹ thuật xử lý mất cân bằng dữ liệu để cải thiện chất lượng dự đoán trong các hệ thống có phân bố lỗi không đồng đều.

- Đề xuất và đánh giá thử nghiệm việc ứng dụng các mô hình sinh dữ liệu đối kháng GAN và các biến thể của chúng nhằm giải quyết vấn đề mất cân bằng dữ liệu trong các tập dữ liệu lỗi phần mềm dạng bảng. Áp dụng các mô hình GAN nhằm tạo dữ liệu tổng hợp chất lượng cao cho lớp thiểu số, từ đó cân bằng phân phối lớp, cải thiện độ chính xác và tăng khả năng khái quát hóa của các mô hình dự đoán lỗi phần mềm. Kết quả nghiên cứu cung cấp cơ sở thử nghiệm quan trọng giúp các nhà nghiên cứu lựa chọn phương pháp sinh dữ liệu phù hợp cho các bài toán học máy có dữ liệu mất cân bằng trong lĩnh vực công nghệ phần mềm. Trên cơ sở tập đặc trưng được tối ưu, luận án đề xuất kết hợp các phương pháp lựa chọn đặc trưng Recursive Feature Elimination (RFE) và Autoencoder với các mô hình sinh dữ liệu dựa trên mạng đối kháng GAN nhằm giải quyết bài toán mất cân bằng dữ liệu trong dự đoán lỗi phần mềm. Kết quả thực nghiệm

cho thấy việc kết hợp RFE và Autoencoder với các biến thể GAN, đặc biệt là CTGAN, WGANGP và VanillaGAN, tạo ra tập dữ liệu huấn luyện có chất lượng cao hơn, giúp các mô hình dự đoán đạt hiệu suất vượt trội so với các kỹ thuật lấy mẫu truyền thống như SMOTE và ADASYN. Kết quả này khẳng định rằng việc tối ưu hóa đồng thời tập đặc trưng và cân bằng dữ liệu là hướng tiếp cận hiệu quả, góp phần nâng cao độ chính xác, khả năng tổng quát hóa và tính ổn định của các mô hình dự đoán lỗi phần mềm.

- Luận án đề xuất giải pháp khai thác thông tin cú pháp và ngữ nghĩa của chương trình để mô hình có thể học được các đặc trưng ngữ nghĩa quan trọng, qua đó nâng cao độ chính xác và khả năng khái quát hóa trong dự đoán lỗi phần mềm. Từ đó, đề xuất mô hình dự đoán lỗi phần mềm dựa trên kiến trúc Transformer, kết hợp với embedding bằng Gensim FastText được trích xuất từ cây AST của mã nguồn có khả năng học các đặc trưng ngữ nghĩa quan trọng của chương trình phần mềm giúp cải thiện đáng kể hiệu quả dự đoán lỗi. Mô hình có thể nắm bắt thông tin ngữ nghĩa trong các biểu diễn vec-tơ của các token mã nguồn, áp dụng cho cả trong cùng dự án và giữa các dự án. Ngoài ra, luận án thực hiện kiểm định Wilcoxon để đánh giá mức độ khác biệt có ý nghĩa thống kê giữa mô hình FastText-Transformer và các mô hình đối sánh. Kết quả đánh giá trên 10 dự án Java cho thấy mô hình được đề xuất của luận án vượt trội hơn so với các mô hình dự đoán lỗi mới hiện nay.

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

PGS.TS. Nguyễn Thanh Bình

TS. Lê Thị Mỹ Hạnh

Hà Thị Minh Phương