

TRÍCH YẾU LUẬN ÁN TIẾN SĨ

Tên luận án: “DỰ ĐOÁN LỖI PHẦN MỀM DỰA TRÊN MÃ NGUỒN”

Tác giả: Hà Thị Minh Phương

Ngành: Khoa học máy tính

Mã số: 9.48.01.01

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

1. Mục tiêu và đối tượng nghiên cứu của luận án

1.1. Mục tiêu của luận án

Mục tiêu của luận án là nghiên cứu về bài toán dự đoán lỗi phần mềm, tập trung vào việc phân tích các vấn đề, hạn chế và thách thức còn tồn tại trong các phương pháp dự đoán hiện nay. Trên cơ sở đó, luận án nghiên cứu, đề xuất và phát triển các kỹ thuật học máy và học sâu nhằm nâng cao hiệu suất của mô hình dự đoán lỗi phần mềm. Luận án tập trung vào mục tiêu cụ thể sau:

- Thứ nhất, luận án nghiên cứu tổng quan về dự đoán lỗi phần mềm và các kỹ thuật được nghiên cứu trong dự đoán lỗi phần mềm; tình hình nghiên cứu và ứng dụng của dự đoán lỗi phần mềm trong nước và trên thế giới.
- Thứ hai, luận án tập trung vào việc 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.
- Thứ ba, tiếp theo, trên cơ sở tập đặc trưng đã được tối ưu, luận án tập trung vào việc giải quyết vấn đề mất cân bằng lớp trong các mô hình dự đoán lỗi phần mềm nhằm nâng cao hiệu suất và độ chính xác của mô hình. Luận án hướng đến việc phân tích ảnh hưởng của sự mất cân bằng dữ liệu đến hiệu năng mô hình dự đoán lỗi, đồng thời đề xuất áp dụng các kỹ thuật lấy mẫu dữ liệu để cân bằng lại tập huấn luyện kết hợp với áp dụng các phương pháp chọn đặc trưng; đặc biệt luận án đề xuất sử dụng mạng đối kháng (GAN) như một giải pháp tạo dữ liệu tổng hợp có tính chân thực cao hơn, phản ánh tốt hơn phân phối toàn cục và mối quan hệ giữa các đặc trưng. Việc kết hợp hai nhóm kỹ thuật lựa chọn đặc trưng và cân bằng dữ liệu hướng tới mục tiêu tối ưu hóa toàn diện dữ liệu đầu vào, qua đó nâng cao hiệu suất và độ chính xác của các mô hình dự đoán lỗi phần mềm.

- Thứ tư, trên cơ sở những giới hạn của các mô hình dự đoán dựa trên đặc trưng được thiết kế thủ công, luận án nghiên cứu hướng tiếp cận dựa trên học sâu nhằm khai thác trực tiếp thông tin cú pháp và ngữ nghĩa của mã nguồn. Cụ thể, 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, luận án 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.

1.2. Đối tượng nghiên cứu của luận án

Nghiên cứu tập trung vào:

- Các phương pháp và kỹ thuật trong dự đoán lỗi phần mềm, các kỹ thuật lựa chọn đặc trưng, xử lý mất cân bằng dữ liệu nhằm tối ưu tập dữ liệu đầu vào.
- Phương pháp trích xuất đặc trưng ngữ nghĩa và cú pháp của mã nguồn.
- Các mô hình word-embedding để học các véc-tơ biểu diễn cho chuỗi token trong mã nguồn.
- Các mô hình học máy và học sâu được sử dụng để xây dựng hệ thống dự đoán lỗi phần mềm có chất lượng.

2. Các phương pháp nghiên cứu đã sử dụng

Phương pháp nghiên cứu bao gồm:

- Phương pháp nghiên cứu tổng quan các công trình liên quan.
- Phương pháp phân tích, tổng hợp các nghiên cứu đã có. Đánh giá các ưu, nhược điểm nghiên cứu liên quan.
- Phương pháp đề xuất giải pháp dựa trên các nghiên cứu đã có và giải quyết các hạn chế.
- Phương pháp thực nghiệm (cài đặt giải pháp đề xuất, đánh giá giải pháp trên các bộ dữ liệu thực nghiệm) và đánh giá, so sánh kết quả thực nghiệm với các nghiên cứu liên quan.

3. Những đóng góp chính

3.1. Đóng góp khoa học

- Đề xuất kết hợp hiệu quả các kỹ thuật chọn lọc và trích xuất đặc trưng với các phương pháp lấy mẫu hiện đại chỉ ra rằng sự phối hợp giữa kỹ thuật chọn lọc đặc trưng RFE, trích xuất đặc trưng bằng Autoencoders, và các biến thể của mô hình GAN (như VanillaGAN, CTGAN, WGANGP, CopulaGAN, TGAN) giúp nâng

cao đáng kể hiệu suất của các mô hình học máy truyền thống trong môi trường dữ liệu mất cân bằng.

- Đề xuất mô hình FastText-Transformer mới dựa trên kiến trúc Transformer kết hợp với biểu diễn từ FastText đã thể hiện khả năng học đặc trưng ngữ nghĩa sâu sắc từ mã nguồn dựa trên cây cú pháp trừu tượng, từ đó cải thiện đáng kể hiệu suất mô hình dự đoán lỗi, đặc biệt trong bối cảnh dự đoán trong cùng và liên dự án.

3.2. Đóng góp thực tiễn

- Đề xuất các phương pháp tối ưu hóa dữ liệu đầu vào và mô hình dự đoán lỗi phần mềm, góp phần nâng cao hiệu quả phát hiện lỗi trong quá trình phát triển và kiểm thử phần mềm.
- Cung cấp cơ sở thực nghiệm cho việc lựa chọn các kỹ thuật lựa chọn đặc trưng, cân bằng dữ liệu và mô hình học sâu phù hợp, hỗ trợ xây dựng các hệ thống dự đoán lỗi có độ chính xác, khả năng khái quát hóa và tính ổn định cao trong các dự án phần mềm thực tế.

4. Kết quả chính

- Luận án đã đề xuất giải pháp phân tích, đánh giá hiệu quả vai trò của các phương pháp lựa chọn đặc trưng trong việc nâng cao chất lượng của mô hình dự đoán lỗi. Đồng thời, đề xuất kết hợp các kỹ thuật xử lý mất cân bằng dữ liệu với các phương pháp chọn đặc trưng để cải thiện hơn nữa chất lượng dự đoán các mô hình.
- Luận án đã đề xuất giải pháp tối ưu và cân bằng dữ liệu đầu vào bằng cách kết hợp các phương pháp lựa chọn đặc trưng (Recursive Feature Elimination và Autoencoders) với các biến thể GAN (bao gồm CTGAN, WGANP, VanillaGAN, CopulaGAN và TGAN) để sinh thêm mẫu cho lớp thiểu số. Kết quả thực nghiệm trên năm bộ dữ liệu lỗi phần mềm cho thấy CTGAN, WGANP và VanillaGAN cho hiệu suất vượt trội so với các phương pháp truyền thống như SMOTE hay ADASYN, đồng thời khẳng định rằng việc kết hợp chọn đặc trưng và sinh dữ liệu bằng GAN giúp nâng cao độ hiệu suất và khả năng khái quát của mô hình dự đoán lỗi phần mềm.
- Luận án đã đề xuất, xây dựng và thử nghiệm thành công mô hình dự đoán lỗi phần mềm FastText-Transformer, kết hợp giữa mô hình nhúng từ Gensim FastText và kiến trúc học sâu Transformer nhằm khai thác hiệu quả thông tin ngữ nghĩa và cú pháp tiềm ẩn trong mã nguồn thông qua cấu trúc cây cú pháp trừu tượng. Từ đó tạo tiền đề cho việc áp dụng mô hình này trong các công cụ hỗ trợ kiểm thử phần mềm và đánh giá chất lượng mã nguồn tự động.

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