HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Trần Thị Hương Giang

ỨNG DỤNG GRAPH NEURAL NETWORK VÀ KỸ THUẬT XỬ LÝ DỮ LIỆU MẤT CÂN BẰNG TRONG PHÁT HIỆN GIAO DỊCH GIAN LẬN TÀI CHÍNH

Chuyên ngành: Khoa học máy tính

Mã số: 8.48.01.01 (Khoa học máy tính)

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

(***Theo định hướng ứng dụng***)

NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS. TS. ĐỖ XUÂN CHỢ

HÀ NỘI - NĂM 2025

LỜI CAM ĐOAN

Tôi cam đoan đây là công trình nghiên cứu của riêng tôi. Các số liệu, kết quả nêu trong đề án tốt nghiệp là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Tác giả đề án tốt nghiệp



Trần Thị Hương Giang

LỜI CẢM ƠN

Trong quá trình thực hiện đề tài, tôi đã nhận được sự hướng dẫn, hỗ trợ và động viên quý báu từ nhiều cá nhân và đơn vị. Tôi xin được bày tỏ lòng biết ơn chân thành đến tất cả những người đã đóng góp để đề tài được hoàn thành.

Trước hết, tôi xin gửi lời cảm ơn sâu sắc đến **giảng viên hướng dẫn**, người đã tận tình định hướng khoa học, góp ý chuyên môn và hỗ trợ tôi trong suốt quá trình xây dựng và hoàn thiện đề tài. Những nhận xét và chỉ dẫn của thầy/cô là nền tảng quan trọng giúp tôi tiếp cận đúng hướng và nâng cao chất lượng nghiên cứu.

Tôi cũng xin trân trọng cảm ơn **các thầy cô trong khoa**, những người đã cung cấp kiến thức nền tảng và truyền cảm hứng trong quá trình học tập và nghiên cứu.

Xin chân thành cảm ơn **gia đình và bạn bè** đã luôn động viên, tạo điều kiện thuận lợi và đồng hành cùng tôi trong suốt thời gian thực hiện đề tài.

Cuối cùng, tôi xin cảm ơn tất cả những ai đã hỗ trợ tôi trực tiếp hoặc gián tiếp. Những đóng góp đó đều có ý nghĩa quan trọng để tôi hoàn thành nghiên cứu này.

Xin trân trọng cảm ơn!

Tác giả đề án tốt nghiệp

Trần Thị Hương Giang

DANH SÁCH HÌNH VẼ

1	Cấu trúc mô hình đề xuất	14
2	Quy trình tiền xử lý dữ liệu	20
3	Confusion matrix của kiến trúc có áp dụng SMOTE trên tập kiểm tra.	40
4	Đường Precision–Recall của hai mô hình trên tập kiểm tra với tỷ lệ gian lận dưới 0.1%	41
5	Đường ROC của hai mô hình trên tập kiểm tra.	42

DANH SÁCH BẢNG

3.1 Bảng so sánh hai mô hình 43
3.2 Tóm tắt so sánh với các nghiên cứu liên quan 46

DANH MỤC CÁC CHỮ VIẾT TẮT

ML	Machine Learning	Học máy
LLM	Large Language Model	Mô hình ngôn ngữ lớn
GNN	Graph Neural Network	Mạng nơ-ron đồ thị
GAT	Graph Attention Network	Mạng nơ-ron đồ thị có cơ chế chú ý
MLP	Multi-Layer Perceptron	Mạng nơ-ron truyền thẳng nhiều tầng
NodeMLP	Mạng MLP dùng để học embedding cho node	Mạng MLP dùng để học biểu diễn cho nút
RF	Random Forest	
SMOTE	Synthetic Minority Over-sampling Technique	Kỹ thuật tăng mẫu tổng hợp cho lớp thiểu số
PR-AUC	Precision-Recall Area Under the Curve	Diện tích dưới đường cong Precision–Recall
ROC-AUC	Receiver Operating Characteristic - Area Under the Curve	Diện tích dưới đường cong Precision–Recall
F1-score	Chỉ số F1 (trung bình điều hòa của Precision và Recall)	Chỉ số F1 (trung bình điều hòa của độ chính xác và độ bao phủ)
Precision	Độ chính xác dương tính	Độ chính xác dương tính
Recall	Độ bao phủ(độ nhạy)	Độ bao phủ (độ nhạy)
K-Means	Thuật toán phân cụm K-Means	Thuật toán phân cụm K-Means
XAI	Explainable Artificial Intelligence	Trí tuệ nhân tạo có khả năng giải thích
GPU	Graphics Processing Unit	Bộ xử lý đồ họa
CSV	Comma-Separated Values	Định dạng dữ liệu phân tách bằng dấu phẩy
PaySim	Payment Simulation Dataset	Bộ dữ liệu mô phỏng giao dịch thanh toán

MỤC LỤC

MỞ ĐẦU 1

1 CƠ SỞ LÝ THUYẾT 5

1.1 Giới thiệu chung 5

1.2 Tổng quan về gian lận giao dịch tài chính 5

1.2.1 Khái niệm gian lận giao dịch tài chính 5

1.2.2 Đặc điểm của dữ liệu giao dịch tài chính 5

1.2.2.1 Quy mô lớn và phát sinh liên tục 5

1.2.2.2 Mất cân bằng dữ liệu nghiêm trọng 6

1.2.2.3 Phụ thuộc quan hệ và theo chuỗi 6

1.3 Cấu trúc dữ liệu giao dịch tài chính 6

1.3.1 Một giao dịch điển hình 6

1.3.2 Giao dịch theo chuỗi và theo tài khoản 7

1.4 Hoạt động bất thường trong giao dịch tài chính 7

1.4.1 Khái niệm hoạt động bất thường 7

1.4.2 Các hình thức gian lận giao dịch phổ biến 7

1.5 Các phương pháp phát hiện gian lận giao dịch 8

1.5.1 Bài toán phát hiện bất thường trong dữ liệu giao dịch 8

1.5.2 Cách tiếp cận dựa trên học máy truyền thống 9

1.5.3 Cách tiếp cận dựa trên học sâu và học biểu diễn 9

1.5.4 Cách tiếp cận dựa trên phát hiện bất thường (Anomaly Detection) 10

1.5.5 Cách tiếp cận khai thác cấu trúc quan hệ và học trên đồ thị 10

1.5.6 Nhận xét chung và định hướng nghiên cứu 11

1.6 Tổng quan các công trình liên quan 11

1.7 Kết luận chương 12

2 PHƯƠNG PHÁP NGHIÊN CỨU VÀ THIẾT KẾ THỰC NGHIỆM 14

2.1 Giới thiệu chung 14

2.2 Dữ liệu 15

2.2.1 Tập dữ liệu 15
2.2.2 Khai phá và phân tích dữ liệu ban đầu 16
2.2.3 Tiền xử lý dữ liệu 16
2.2.3.1 Làm sạch dữ liệu và xử lý giá trị thiếu 17
2.2.3.2 Chuẩn hóa thang đo và xử lý phân bố lệch 17
2.2.3.3 Xây dựng đặc trưng thời gian (Temporal Encoding) 18
2.2.3.4 Phân chia dữ liệu theo thứ tự thời gian 19
2.2.3.5 Vai trò của tiền xử lý trong cấu trúc tổng thể 20
2.2.4 Tóm tắt mô hình tiền xử lý dữ liệu 20
2.2.5 Định dạng dữ liệu sau tiền xử lý 21
2.3 Biểu diễn dữ liệu và xây dựng đồ thị 22
2.3.1 Gom cụm tài khoản (Account Clustering) 22
2.3.2 Mô hình gom cụm tài khoản bằng K-Means 24
2.3.3 Xây dựng đồ thị giao dịch theo cụm 26
2.3.4 Vai trò của biểu diễn đồ thị trong mô hình tổng thể 27
2.4 Học biểu diễn và mô hình phân loại 28
2.4.1 Học embedding bằng NodeMLP 28
2.4.2 Mô hình hóa quan hệ bằng Graph Attention Network 29
2.4.2.1 Cơ chế attention 30
2.4.2.2 Cập nhật embedding node 31
2.4.3 Edge embedding và biểu diễn giao dịch 31
2.4.4 Xử lý mất cân bằng dữ liệu bằng SMOTE 32
2.4.5 Bộ phân loại Random Forest 33
2.5 Thiết kế huấn luyện và phân tích 34
2.5.1 Pipeline huấn luyện 34
2.5.2 Các thước đo đánh giá 34
2.6 Kết luận chương 34
3 TRIỂN KHAI THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ 36
3.1 Giới thiệu chung 36

3.2 Môi trường và công cụ thực nghiệm 36
3.3 Thiết lập thực nghiệm 37
3.3.1 Dữ liệu và phân chia tập 37
3.3.2 Mô hình không áp dụng SMOTE (Baseline) 38
3.3.3 Mô hình chính (có áp dụng SMOTE) 38
3.3.4 Chiến lược huấn luyện và chọn ngưỡng 39
3.4 Kết quả thực nghiệm và phân tích 39
3.4.1 Kết quả mô hình baseline 39
3.4.2 Mô hình áp dụng SMOTE (Nhánh chính) 40
3.4.3 So sánh hai mô hình 41
3.5 So sánh với các nghiên cứu liên quan 43
3.6 Kết luận chương 47
KẾT LUẬN **49**
TÀI LIỆU THAM KHẢO **52**
PHỤ LỤC **56**
PHỤ LỤC A. MÔ TẢ DỮ LIỆU VÀ TIỀN XỬ LÝ 56
PHỤ LỤC B. THIẾT LẬP PIPELINE VÀ THAM SỐ MÔ HÌNH 58
PHỤ LỤC C. MINH HOẠ VÀ BẢNG BIỂU BỔ SUNG 60
PHỤ LỤC D. THIẾT LẬP THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ 61

MỞ ĐẦU

Lý do chọn đề tài

Trong các hệ thống tài chính – ngân hàng hiện đại, khối lượng giao dịch điện tử phát sinh ngày càng lớn, đa dạng về loại hình và được xử lý với tốc độ rất cao. Cùng với sự phát triển đó, các hành vi gian lận tài chính cũng trở nên phức tạp hơn, không chỉ xuất hiện dưới dạng các giao dịch bất thường riêng lẻ mà thường có mối liên hệ giữa nhiều tài khoản, nhiều giao dịch và diễn ra theo chuỗi hoặc theo mạng lưới trung gian.

Các phương pháp nhận diện gian lận truyền thống chủ yếu dựa trên dữ liệu dạng bảng và các thuật toán ML (Machine Learning) cơ bản. Hướng tiếp cận này có thể đạt hiệu quả nhất định khi các giao dịch được mô tả đầy đủ bằng các đặc trưng độc lập. Tuy nhiên, khi dữ liệu giao dịch có cấu trúc phức tạp, nhiều chiều và tồn tại các mối quan hệ giữa các thực thể như tài khoản, giao dịch và dòng tiền, các mô hình này gặp hạn chế do không khai thác trực tiếp được các thông tin liên quan.

Bên cạnh đó, dữ liệu gian lận tài chính có đặc điểm mất cân bằng rất nghiêm trọng, khi số lượng giao dịch gian lận chỉ chiếm một tỷ lệ rất nhỏ so với tổng số giao dịch hợp lệ. Điều này khiến nhiều mô hình có thể đạt kết quả tốt theo các thước đo tổng quát, nhưng lại không phát hiện được các trường hợp gian lận hiếm gặp trong thực tế vận hành.

Trong bối cảnh đó, việc nghiên cứu và ứng dụng các phương pháp học theo cấu trúc đồ thị, đặc biệt là GNN, kết hợp với các kỹ thuật xử lý mất cân bằng dữ liệu, được xem là một hướng tiếp cận phù hợp nhằm nâng cao hiệu quả phát hiện gian lận giao dịch tài chính.

Mục tiêu nghiên cứu

Mục tiêu của đề án là xây dựng và đánh giá một pipeline phát hiện giao dịch gian lận tài chính dựa trên học theo cấu trúc đồ thị, trong đó trọng tâm là khai thác khả năng học biểu diễn của GNN kết hợp với các kỹ thuật xử lý mất cân bằng dữ

liệu, nhằm cải thiện khả năng phát hiện các giao dịch gian lận hiếm gặp trong dữ liệu giao dịch quy mô lớn.

Các mục tiêu cụ thể của đề án bao gồm:

- Xây dựng mô hình phát hiện gian lận bao gồm các bước tiền xử lý dữ liệu, biểu diễn giao dịch dưới dạng đồ thị, học nhúng bằng GNN và phân loại giao dịch.
- Triển khai thực nghiệm hai hướng tiếp cận: mô hình không áp dụng kỹ thuật cân bằng dữ liệu và mô hình có áp dụng SMOTE, nhằm đánh giá tác động của việc xử lý mất cân bằng dữ liệu.
- Đánh giá hiệu quả của cấu trúc này thông qua các thước đo phù hợp với dữ liệu mất cân bằng, đặc biệt là PR-AUC, F1-score, Precision và Recall.
- Phân tích khả năng ứng dụng của mô hình trong bối cảnh giám sát giao dịch tài chính thực tế.

Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu của đề án là bài toán phát hiện giao dịch gian lận tài chính trên dữ liệu giao dịch điện tử, trong đó các giao dịch và tài khoản được mô hình hóa dưới dạng đồ thị.

Phạm vi nghiên cứu của đề án được giới hạn như sau:

- Dữ liệu sử dụng là dữ liệu giao dịch tài chính công khai hoặc dữ liệu mô phỏng, phản ánh đặc trưng mất cân bằng mạnh của nhãn gian lận.
- Mô hình tập trung vào dữ liệu có cấu trúc, chưa xem xét dữ liệu phi cấu trúc như văn bản mô tả giao dịch hoặc ghi chú nghiệp vụ.
- Các mô hình ngôn ngữ lớn và các kỹ thuật embedding nâng cao là hướng mở rộng và không nằm trong phạm vi thực nghiệm chính của đề án.

Phương pháp nghiên cứu

Đề án sử dụng phương pháp nghiên cứu kết hợp giữa lý thuyết và thực nghiệm, bao gồm các bước chính sau:

- Nghiên cứu và tổng hợp cơ sở lý thuyết liên quan đến gian lận tài chính, dữ liệu giao dịch và các phương pháp học máy, học sâu trên dữ liệu đồ thị.
- Tiền xử lý dữ liệu giao dịch, xây dựng đặc trưng và biểu diễn dữ liệu dưới dạng đồ thị thông qua việc gom cụm tài khoản.
- Áp dụng GNN để học biểu diễn quan hệ giữa các cụm tài khoản.
- Kết hợp kỹ thuật xử lý mất cân bằng dữ liệu và bộ phân loại truyền thống để phát hiện giao dịch gian lận trên bộ dữ liệu PaySim.
- Đánh giá và phân tích kết quả thực nghiệm dựa trên các thước đo phù hợp với dữ liệu mất cân bằng thông qua 2 mô hình có sử dụng/ không sử dụng kỹ thuật xử lý mất cân bằng thông qua các chỉ số như ROC-AUC, PR-AUC, F1-score.

Đóng góp chính của đề án

Các đóng góp chính của đề án bao gồm:

- Đề xuất một cấu trúc mô hình tích hợp kết hợp biểu diễn đồ thị, học sâu và học máy truyền thống cho bài toán phát hiện giao dịch gian lận tài chính.
- Làm rõ vai trò của GNN trong việc khai thác cấu trúc quan hệ giữa các giao dịch và tài khoản.
- Đánh giá tác động của kỹ thuật cân bằng dữ liệu SMOTE trong không gian embedding đối với hiệu quả phát hiện gian lận hiếm gặp.
- Cung cấp một hướng tiếp cận có khả năng mở rộng và phù hợp để áp dụng trong các hệ thống giám sát giao dịch tài chính.

Bố cục của đề án

Nội dung của đề án được tổ chức như sau:

- **Chương 1 - CƠ SỞ LÝ THUYẾT**: Trình bày tổng quan về gian lận tài chính, đặc điểm và cấu trúc dữ liệu giao dịch, cũng như các khái niệm và thuật toán nền tảng được sử dụng trong đề án.

- **Chương 2 - PHƯƠNG PHÁP NGHIÊN CỨU VÀ THIẾT KẾ THỰC NGHIỆM** : Trình bày quy trình xây dựng pipeline phát hiện gian lận, bao gồm khảo sát dữ liệu, tiền xử lý, biểu diễn đồ thị và thiết kế mô hình.

- **Chương 3 - TRIỂN KHAI THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ**: Mô tả thiết lập thực nghiệm, các nhánh đánh giá và các thước đo được sử dụng. Trình bày kết quả thực nghiệm, phân tích và so sánh các hướng tiếp cận.

- **Chương KẾT LUẬN**: Thảo luận kết quả, hạn chế của đề án và đề xuất các hướng mở rộng trong tương lai.

CHƯƠNG 1: CƠ SỞ LÝ THUYẾT

1.1 Giới thiệu chung

Sau đây tôi xin trình bày cơ sở lý thuyết và khung khái niệm phục vụ cho bài toán phát hiện giao dịch gian lận tài chính. Nội dung chương tập trung làm rõ bối cảnh dữ liệu giao dịch trong các hệ thống tài chính hiện đại, đặc điểm và cấu trúc của giao dịch tài chính, cũng như các khái niệm và phương pháp nền tảng thường được sử dụng trong nghiên cứu phát hiện gian lận.

Mục tiêu của chương là chuẩn hóa hệ thuật ngữ, xác lập phạm vi bài toán và cung cấp nền tảng lý thuyết cho việc nghiên cứu các phương pháp phát hiện gian lận ở các chương tiếp theo. Trên cơ sở đó, các hướng tiếp cận dựa trên học máy, học sâu và học theo cấu trúc đồ thị được giới thiệu ở mức tổng quan nhằm làm rõ bối cảnh nghiên cứu và các xu hướng phương pháp hiện nay.

1.2 Tổng quan về gian lận giao dịch tài chính

1.2.1 Khái niệm gian lận giao dịch tài chính

Gian lận giao dịch tài chính có thể được hiểu là các hành vi cố ý thực hiện giao dịch nhằm chiếm đoạt tài sản, làm sai lệch trạng thái tài chính hoặc khai thác các điểm yếu trong quy trình xử lý giao dịch của hệ thống. Các hành vi này có thể xuất hiện trong nhiều hình thức khác nhau như chuyển tiền trái phép, lợi dụng hệ thống thanh toán điện tử, gian lận thẻ, hoặc thao túng chuỗi giao dịch để che giấu nguồn gốc tài sản.

Trong thực tế vận hành, gian lận không phải lúc nào cũng thể hiện rõ ràng ở từng giao dịch riêng lẻ. Nhiều giao dịch gian lận chỉ bộc lộ dấu hiệu bất thường khi được xem xét trong mối quan hệ với các giao dịch khác, với lịch sử hoạt động của tài khoản hoặc trong một khoảng thời gian nhất định.

1.2.2 Đặc điểm của dữ liệu giao dịch tài chính

1.2.2.1 Quy mô lớn và phát sinh liên tục

Dữ liệu giao dịch tài chính thường có quy mô rất lớn, được sinh ra liên tục theo thời gian và yêu cầu xử lý gần thời gian thực. Điều này đặt ra thách thức không nhỏ đối với các hệ thống phát hiện gian lận, đặc biệt khi cần phân tích hành vi trong thời gian ngắn.

1.2.2.2 Mất cân bằng dữ liệu nghiêm trọng

Một đặc điểm nổi bật của dữ liệu gian lận là sự mất cân bằng cực lớn giữa các lớp. Trong hầu hết các tập dữ liệu thực tế, số lượng giao dịch gian lận chỉ chiếm một tỷ lệ rất nhỏ so với tổng số giao dịch hợp lệ. Điều này khiến nhiều mô hình có xu hướng thiên lệch về lớp đa số và bỏ sót các trường hợp gian lận hiếm gặp.

1.2.2.3 Phụ thuộc quan hệ và theo chuỗi

Giao dịch tài chính không tồn tại độc lập mà thường có mối liên hệ với các giao dịch khác thông qua tài khoản gửi, tài khoản nhận hoặc thông qua chuỗi hành vi theo thời gian. Các mẫu gian lận phức tạp thường xuất hiện dưới dạng chuỗi giao dịch liên tiếp hoặc theo nhóm tài khoản, thay vì một giao dịch đơn lẻ.

Tuy nhiên, đối với một số bộ dữ liệu mô phỏng thường được sử dụng trong nghiên cứu, điển hình như PaySim, mức độ liên kết giữa các giao dịch và tài khoản có thể không phản ánh đầy đủ độ phức tạp của dữ liệu thực tế. Điều này đặt ra yêu cầu cần có các phương pháp biểu diễn và xử lý phù hợp khi áp dụng các mô hình học theo cấu trúc quan hệ.

1.3 Cấu trúc dữ liệu giao dịch tài chính

1.3.1 Một giao dịch điển hình

Một giao dịch tài chính điển hình trong hệ thống thanh toán điện tử thường bao gồm các thành phần chính sau:

- Tài khoản gửi và tài khoản nhận.
- Giá trị giao dịch.
- Thời điểm phát sinh giao dịch.
- Loại giao dịch (chuyển tiền, thanh toán, rút tiền, v.v.).
- Trạng thái giao dịch, các thông tin liên quan đến số dư trước và sau giao

dịch.

Các thành phần này cung cấp cơ sở để trích xuất đặc trưng phục vụ cho các mô hình phát hiện gian lận.

1.3.2 Giao dịch theo chuỗi và theo tài khoản

Ngoài các thuộc tính cục bộ, giao dịch còn mang thông tin ngữ cảnh quan trọng khi được xem xét theo lịch sử hoạt động của tài khoản. Các đặc trưng như tần suất giao dịch, tổng giá trị giao dịch trong một khoảng thời gian, hoặc mối quan hệ giữa các tài khoản là những yếu tố then chốt giúp nhận diện hành vi bất thường.

1.4 Hoạt động bất thường trong giao dịch tài chính

1.4.1 Khái niệm hoạt động bất thường

Hoạt động bất thường được hiểu là các giao dịch hoặc chuỗi giao dịch có đặc điểm lệch đáng kể so với mô hình hành vi thông thường của hệ thống. Cần lưu ý rằng không phải mọi bất thường đều là gian lận; tuy nhiên, các giao dịch gian lận thường nằm trong tập con của các hoạt động bất thường có tính chủ ý.

Hoạt động bất thường có thể xuất hiện dưới nhiều dạng:

- Bất thường điểm: giao dịch riêng lẻ có giá trị hoặc thời điểm bất thường.
- Bất thường theo ngữ cảnh: giao dịch bất thường trong bối cảnh cụ thể.
- Bất thường tập thể: một nhóm giao dịch cấu thành hành vi bất thường.
- Thay đổi chế độ: sự thay đổi kéo dài trong chuỗi giao dịch của tài khoản.

1.4.2 Các hình thức gian lận giao dịch phổ biến

Trong hệ thống tài chính, gian lận có thể xuất hiện dưới nhiều hình thức như:

- Gian lận chuyển tiền thông qua tài khoản trung gian.
- Chia nhỏ giao dịch để tránh bị phát hiện.
- Lợi dụng thời gian ít giám sát để thực hiện chuỗi giao dịch bất thường.

- Thao túng hành vi giao dịch để đánh lạc hướng hệ thống giám sát.

Các hình thức này thường để lại dấu vết dưới dạng chuỗi giao dịch liên quan, mức độ giao dịch tăng đột biến hoặc mối quan hệ bất thường giữa các tài khoản.

1.5 Các phương pháp phát hiện gian lận giao dịch

1.5.1 Bài toán phát hiện bất thường trong dữ liệu giao dịch

Bài toán phát hiện gian lận giao dịch tài chính thường được tiếp cận dưới góc nhìn của bài toán phát hiện bất thường (anomaly detection), trong đó các giao dịch gian lận được xem là những quan sát hiếm và có đặc điểm lệch đáng kể so với hành vi giao dịch thông thường của hệ thống. Trong bối cảnh này, mục tiêu của hệ thống không chỉ là phân loại nhị phân gian lận – không gian lận, mà còn là gán điểm rủi ro cho mỗi giao dịch hoặc mỗi thực thể liên quan (ví dụ: tài khoản, chuỗi giao dịch), từ đó hỗ trợ quá trình xếp hạng và ra quyết định theo một ngưỡng phù hợp.

Một đặc trưng quan trọng của bài toán phát hiện bất thường trong dữ liệu giao dịch là sự khan hiếm nhãn gian lận. Trong nhiều hệ thống tài chính thực tế, việc gán nhãn gian lận đòi hỏi quy trình điều tra thủ công phức tạp, dẫn đến tập dữ liệu huấn luyện thường có tỷ lệ gian lận rất thấp và không phản ánh đầy đủ các dạng gian lận tiềm ẩn. Do đó, nhiều phương pháp tập trung vào việc học mô hình hành vi bình thường của hệ thống và phát hiện các sai lệch so với mô hình này, thay vì học trực tiếp ranh giới phân lớp giữa gian lận và không gian lận.

Tuy nhiên, không phải mọi giao dịch bất thường đều tương ứng với hành vi gian lận. Một số giao dịch hợp lệ vẫn có thể mang đặc điểm bất thường do thay đổi hành vi người dùng, yếu tố thời vụ hoặc các sự kiện ngoại sinh. Điều này khiến bài toán phát hiện bất thường trong dữ liệu giao dịch tài chính trở nên phức tạp hơn so với các bài toán anomaly detection truyền thống, đòi hỏi mô hình không chỉ nhạy với các sai lệch hiếm gặp mà còn phải có khả năng phân biệt giữa bất thường mang tính ngẫu nhiên và bất thường mang tính chủ ý.

Các nghiên cứu gần đây cho thấy rằng bài toán phát hiện gian lận giao dịch thường không thể được giải quyết hiệu quả bằng một phương pháp đơn lẻ. Thay vào đó, các hướng tiếp cận hiện đại thường kết hợp nhiều thành phần khác nhau, bao gồm học biểu diễn để làm giàu không gian đặc trưng, khai thác cấu trúc quan hệ giữa các thực thể giao dịch, cũng như các phương pháp học máy có giám sát trong bối cảnh dữ liệu mất cân bằng nghiêm trọng. Cách tiếp cận đa thành phần này tạo tiền đề cho sự hình thành của các pipeline phát hiện gian lận nhiều tầng, trong đó mỗi nhóm phương pháp đóng vai trò bổ trợ lẫn nhau.

1.5.2 Cách tiếp cận dựa trên học máy truyền thống

Các phương pháp học máy truyền thống là một trong những hướng tiếp cận sớm và phổ biến nhất trong bài toán phát hiện gian lận giao dịch tài chính. Các mô hình như Logistic Regression, Decision Tree, Random Forest hay Gradient Boosting thường được huấn luyện trên dữ liệu dạng bảng, trong đó mỗi giao dịch được biểu diễn bằng một tập đặc trưng thủ công (ví dụ: giá trị giao dịch, tần suất giao dịch, số dư tài khoản, thời gian giao dịch).

Ưu điểm của nhóm phương pháp này là tính đơn giản, dễ triển khai và khả năng giải thích tương đối tốt. Đặc biệt, các mô hình ensemble dựa trên cây quyết định thường cho hiệu quả ổn định trong bối cảnh dữ liệu mất cân bằng khi được kết hợp với các kỹ thuật điều chỉnh trọng số hoặc oversampling. Tuy nhiên, hạn chế lớn của các phương pháp này là phụ thuộc mạnh vào thiết kế đặc trưng thủ công và khả năng khai thác hạn chế các mối quan hệ phức tạp giữa các giao dịch và tài khoản.

1.5.3 Cách tiếp cận dựa trên học sâu và học biểu diễn

Với sự phát triển của học sâu, nhiều nghiên cứu đã áp dụng các mô hình mạng nơ-ron nhân tạo nhằm tự động học biểu diễn dữ liệu giao dịch từ các đặc trưng ban đầu. Các kiến trúc như mạng nơ-ron đa tầng (MLP), mạng hồi tiếp (RNN) hoặc các mô hình dựa trên Transformer cho phép mô hình hóa các quan hệ phi tuyến và các mẫu hành vi theo thời gian.

Cách tiếp cận này giúp giảm phụ thuộc vào thiết kế đặc trưng thủ công, đồng thời cho phép học các biểu diễn liên tục (embedding) giàu thông tin hơn so với dữ liệu bảng truyền thống. Tuy nhiên, trong nhiều trường hợp, các mô hình học sâu thuần túy vẫn gặp khó khăn khi dữ liệu có tính mất cân bằng nghiêm trọng hoặc khi các mối quan hệ giữa giao dịch không được biểu diễn tường minh trong dữ liệu đầu vào.

1.5.4 Cách tiếp cận dựa trên phát hiện bất thường (Anomaly Detection)

Một hướng tiếp cận khác là xem phát hiện gian lận như một bài toán phát hiện bất thường, trong đó các giao dịch gian lận được coi là các điểm dữ liệu hiếm và lệch khỏi phân bố hành vi thông thường. Các phương pháp như Isolation Forest, One-Class SVM hoặc Autoencoder thường được sử dụng trong bối cảnh này.

Ưu điểm của hướng tiếp cận phát hiện bất thường là không yêu cầu hoặc chỉ yêu cầu rất ít dữ liệu có nhãn, phù hợp với các hệ thống thực tế nơi việc gán nhãn gian lận gặp nhiều khó khăn. Tuy nhiên, không phải mọi giao dịch bất thường đều là gian lận, do đó các phương pháp này thường gặp thách thức về tỷ lệ cảnh báo sai cao và khó đáp ứng yêu cầu vận hành trong các hệ thống tài chính quy mô lớn.

1.5.5 Cách tiếp cận khai thác cấu trúc quan hệ và học trên đồ thị

Gần đây, nhiều nghiên cứu đã nhấn mạnh vai trò của việc khai thác cấu trúc quan hệ giữa các thực thể giao dịch, chẳng hạn như mối liên hệ giữa tài khoản gửi, tài khoản nhận và chuỗi giao dịch theo thời gian. Trong cách tiếp cận này, dữ liệu giao dịch được biểu diễn dưới dạng đồ thị, cho phép mô hình hóa trực tiếp các mối quan hệ phức tạp mà dữ liệu bảng truyền thống khó thể hiện.

Các mô hình học trên đồ thị, đặc biệt là mạng nơ-ron đồ thị, cho phép học biểu diễn của các nút và cạnh thông qua cơ chế lan truyền thông tin trên đồ thị. Hướng tiếp cận này đã cho thấy hiệu quả rõ rệt trong việc phát hiện các mẫu gian lận mang tính liên kết, nơi hành vi gian lận chỉ trở nên rõ ràng khi xét trong bối cảnh quan hệ. Tuy nhiên, việc xây dựng đồ thị phù hợp và xử lý các vấn đề như dữ

liệu thưa hoặc mất cân bằng vẫn là những thách thức cần được giải quyết.

1.5.6 Nhận xét chung và định hướng nghiên cứu

Từ các cách tiếp cận trên có thể thấy rằng mỗi nhóm phương pháp đều có những ưu điểm và hạn chế riêng. Các mô hình học máy truyền thống cho khả năng phân loại ổn định, học sâu và học biểu diễn giúp làm giàu không gian đặc trưng, trong khi các mô hình đồ thị cho phép khai thác hiệu quả cấu trúc quan hệ phức tạp giữa các giao dịch.

Do đó, xu hướng nghiên cứu hiện nay hướng tới việc kết hợp các hướng tiếp cận này trong những khuôn khổ tổng quát, nhằm tận dụng ưu điểm của từng phương pháp và khắc phục các hạn chế riêng lẻ.

1.6 Tổng quan các công trình liên quan

Các công trình nghiên cứu liên quan đến phát hiện gian lận giao dịch tài chính có thể được phân nhóm theo ba hướng chính: (i) các phương pháp học máy truyền thống dựa trên dữ liệu bảng, (ii) các hướng tiếp cận sử dụng embedding để làm giàu biểu diễn dữ liệu, và (iii) các mô hình khai thác cấu trúc quan hệ, đặc biệt là các mô hình học trên đồ thị.

Ví dụ, nghiên cứu [1] áp dụng các kỹ thuật học máy trên nền tảng SAS Viya cho bài toán phát hiện gian lận giao dịch, trong đó nhấn mạnh quy trình lựa chọn pipeline phù hợp và xem xét các hướng ensemble cũng như giảm chiều dữ liệu. Trên các thiết lập thực nghiệm, các mô hình theo hướng ensemble/tree-based như RF thường cho kết quả ổn định hơn so với các mô hình tuyến tính trong bối cảnh dữ liệu có quan hệ phi tuyến và phân bố mất cân bằng. Các kết quả này cho thấy các mô hình ensemble/tree-based như Random Forest thường đạt hiệu quả ổn định trong bối cảnh dữ liệu mất cân bằng.

Một số dự án khác tập trung vào mạng nơ-ron nhân tạo. Ví dụ, nghiên cứu [2] sử dụng ANN đạt accuracy rất cao, nhưng do không đồng thời đánh giá

precision và recall trên tập dữ liệu có độ lệch lớn, nên kết quả có thể mang tính lạc quan. So sánh giữa ANN huấn luyện bằng Simulated Annealing (SA-ANN) và mô hình học trực tuyến HTM-CLA cho thấy SA-ANN cho kết quả tốt hơn, trong khi HTM-CLA chưa thực sự vượt trội [3]. Các kết quả này cho thấy mặc dù ANN có thể đạt độ chính xác cao trong một số thiết lập, việc thiếu các thước đo phù hợp cho dữ liệu mất cân bằng và hạn chế trong khai thác cấu trúc quan hệ khiến ANN thuần túy khó đáp ứng yêu cầu vận hành thực tế.

Nghiên cứu [4] hướng đến phát hiện báo cáo tài chính gian lận bằng cách kết hợp các tỷ số tài chính với đặc trưng văn bản (MD&A) sử dụng mạng chú ý phân cấp (HAN). Kết quả cho thấy việc tích hợp đặc trưng văn bản giúp tăng cường sức mạnh phân loại so với chỉ sử dụng dữ liệu tài chính.

Song song đó, nhiều công trình gần đây nhấn mạnh hiệu quả của embedding và graph-based learning. Một số nghiên cứu áp dụng knowledge graph hoặc thực hiện embedding kết hợp với các bộ phân loại như SVM, KNN, Logistic Regression và cho thấy cải thiện đáng kể so với mô hình thuần ML [5], [6]. Ngoài ra, các phương pháp tree-based hiện đại như Random Forest và Gradient Boosting khi kết hợp với embedding thường đạt hiệu quả cao trên dữ liệu mất cân bằng [1], [7], [8], [9]. Những kết quả này cho thấy embedding và học theo cấu trúc đồ thị đóng vai trò then chốt trong việc phát hiện các hành vi gian lận mang tính liên kết.

Từ tổng quan các công trình liên quan có thể thấy rằng mỗi nhóm phương pháp đều giải quyết một khía cạnh riêng của bài toán phát hiện gian lận: học máy truyền thống cho khả năng phân loại ổn định, embedding giúp làm giàu biểu diễn dữ liệu, trong khi các mô hình đồ thị cho phép khai thác cấu trúc quan hệ phức tạp. Tuy nhiên, chưa có nhiều nghiên cứu kết hợp một cách hệ thống các thành phần này trong cùng một pipeline, đặc biệt trong bối cảnh dữ liệu giao dịch có tỷ lệ gian lận cực kỳ thấp.

1.7 Kết luận chương

Tóm lại, chương I đã trình bày cơ sở lý thuyết cho bài toán phát hiện giao dịch gian lận tài chính, bao gồm bối cảnh dữ liệu giao dịch, cấu trúc và đặc điểm của giao dịch tài chính, các khái niệm về hoạt động bất thường, cũng như các hướng tiếp cận phổ biến trong phát hiện gian lận. Thông qua việc tổng quan các phương pháp học máy, học sâu và học theo cấu trúc đồ thị, chương đã làm rõ những thách thức cốt lõi của bài toán, đặc biệt là tính mất cân bằng dữ liệu và mối quan hệ phức tạp giữa các giao dịch.

CHƯƠNG 2: PHƯƠNG PHÁP NGHIÊN CỨU VÀ THIẾT KẾ THỰC NGHIỆM

2.1 Giới thiệu chung

Trong phạm vi đề án, luận văn đề xuất một mô hình phát hiện giao dịch gian lận tài chính dựa trên việc kết hợp học biểu diễn, khai thác cấu trúc quan hệ và các kỹ thuật xử lý dữ liệu mất cân bằng. Khác với các phương pháp tiếp cận đơn lẻ, mô hình được xây dựng theo kiến trúc nhiều tầng, trong đó mỗi thành phần đảm nhiệm một vai trò cụ thể nhằm làm giàu thông tin và cải thiện khả năng phát hiện các giao dịch gian lận hiếm gặp.



Hình 2.1: Cấu trúc mô hình đề xuất

Pipeline đề xuất bao gồm các bước chính sau:

- **Tiền xử lý và trích xuất đặc trưng giao dịch:** Dữ liệu giao dịch thô được làm sạch và chuẩn hóa, sau đó trích xuất các đặc trưng cục bộ của giao dịch và đặc trưng tổng hợp theo tài khoản.

- **Xây dựng đặc trưng và gom cụm tài khoản**

  Các tài khoản được biểu diễn thông qua các đặc trưng hành vi và được gom cụm để giảm độ thưa đồ thị và làm nổi bật các mẫu hành vi ở mức khái quát.

- **Học embedding bằng mạng nơ-ron đa tầng (MLP)**

MLP được sử dụng để ánh xạ dữ liệu bảng sang không gian embedding liên tục, giúp chuẩn hóa và làm giàu biểu diễn trước khi đưa vào mô hình đồ thị.

- **Xây dựng đồ thị giao dịch và học biểu diễn quan hệ bằng GNN**

Đồ thị giao dịch được xây dựng trên các cụm tài khoản, trong đó các cạnh biểu diễn luồng giao dịch. Mô hình GNN, cụ thể là Graph Attention Network (GAT), được sử dụng để học biểu diễn node dựa trên cả đặc trưng nội tại và cấu trúc quan hệ.

- **Xử lý dữ liệu mất cân bằng trong không gian embedding**

Sau khi học được embedding giàu thông tin, các kỹ thuật cân bằng dữ liệu được áp dụng để cải thiện khả năng học ranh giới phân lớp của bộ phân loại.

- **Phân loại gian lận**: Cuối cùng, embedding giao dịch được đưa vào bộ phân loại để dự đoán gian lận, cho phép gán nhãn hoặc điểm rủi ro cho từng giao dịch.

2.2 Dữ liệu

2.2.1 Tập dữ liệu

Thực nghiệm được tiến hành trên tập dữ liệu giao dịch tài chính PaySim, công bố trên nền tảng Kaggle. Trong pipeline đề xuất, dữ liệu đầu vào được lấy từ tập PaySim - một tập dữ liệu mô phỏng nhật ký giao dịch của hệ thống thanh toán di động. Việc sử dụng PaySim cho phép kiểm soát đầy đủ nhãn gian lận, đồng thời tái hiện các đặc điểm cốt lõi của dữ liệu giao dịch thực tế như phân bố lệch và tỷ lệ gian lận cực thấp. Tập dữ liệu bao gồm hơn 6,3 triệu bản ghi, với các thuộc tính chính như tài khoản gửi, tài khoản nhận, giá trị giao dịch, số dư trước và sau giao dịch, loại giao dịch và nhãn gian lận.

Một đặc điểm quan trọng của tập dữ liệu PaySim là tỷ lệ giao dịch gian lận chiếm dưới 0.1% tổng số giao dịch - giống tỷ lệ ngoài đời thực, nhưng với số lượng

tài khoản rất lớn thì phần lớn tài khoản lại chỉ xuất hiện với tần suất giao dịch thấp.

Do đó, dữ liệu PaySim không chỉ đặt ra thách thức về mất cân bằng nhãn, mà còn đặt ra thách thức về mức độ kết nối của đồ thị. Đây là cơ sở quan trọng cho việc lựa chọn chiến lược gom cụm tài khoản ở các bước tiếp theo. Vì lý do đó, các thí nghiệm trong đề tài không chỉ tập trung vào khả năng phân biệt tổng quát mà đặc biệt chú trọng đến hiệu quả nhận diện các mẫu gian lận hiếm gặp.

2.2.2 Khai phá và phân tích dữ liệu ban đầu

Quá trình khảo sát dữ liệu ban đầu tập trung vào việc đánh giá phân bố nhãn, tần suất xuất hiện của tài khoản và các thuộc tính giao dịch chính. Kết quả cho thấy phân bố giá trị giao dịch có dạng nặng đuôi (heavy-tailed), với một số lượng nhỏ giao dịch có giá trị rất lớn trong khi phần lớn giao dịch có giá trị nhỏ. Bên cạnh đó, phân bố số lần xuất hiện của tài khoản cho thấy hiện tượng "đuôi dài", trong đó phần lớn tài khoản chỉ tham gia một hoặc vài giao dịch. Điều này làm cho việc xây dựng đồ thị ở mức tài khoản đơn lẻ trở nên kém hiệu quả và củng cố lựa chọn gom cụm tài khoản trước khi áp dụng GNN. Các quan sát này trực tiếp định hướng các bước tiền xử lý tiếp theo, bao gồm biến đổi log cho các biến có phân bố lệch mạnh và chiến lược gom cụm để cải thiện tính kết nối của đồ thị.

2.2.3 Tiền xử lý dữ liệu

Tiền xử lý dữ liệu được hiểu là tập hợp các phép biến đổi có kiểm soát nhằm chuyển dữ liệu giao dịch thô thành dạng đặc trưng đầu vào phù hợp cho các bước học biểu diễn và phân loại trong pipeline. Đầu vào của bước này là các bản ghi giao dịch cùng các thuộc tính gốc; đầu ra là bộ đặc trưng đã được làm sạch, chuẩn hóa thang đo và mã hóa thời gian, bảo đảm tính nhất quán giữa các tập train/validation/test. Trên cơ sở đó, các bước tiếp theo (gom cụm, xây dựng đồ thị và học biểu diễn) có thể sử dụng dữ liệu ở dạng ổn định và tránh rò rỉ thông tin theo thời gian.

Mục tiêu của bước tiền xử lý dữ liệu là chuyển đổi dữ liệu giao dịch thô thành dạng đầu vào phù hợp cho pipeline học biểu diễn và phân loại, đồng thời bảo toàn các tín hiệu hành vi quan trọng phục vụ phát hiện gian lận. Quá trình tiền xử lý trong đề án được thiết kế nhằm đạt ba mục tiêu chính: (i) giảm nhiễu kỹ thuật và ảnh hưởng của các giá trị ngoại lai thường gặp trong dữ liệu tài chính, (ii) hạn chế rủi ro rò rỉ thông tin từ tương lai vào quá trình huấn luyện, và (iii) chuẩn hóa dữ liệu để các mô hình học máy và học sâu có thể hội tụ ổn định và so sánh công bằng.

Quy trình tiền xử lý dữ liệu trong đề án được thực hiện theo các bước sau:

2.2.3.1 Làm sạch dữ liệu và xử lý giá trị thiếu

Tập dữ liệu ban đầu có thể chứa các bản ghi không hợp lệ hoặc các giá trị thiếu do đặc thù quá trình mô phỏng và tổng hợp dữ liệu. Các bản ghi thiếu thông tin cốt lõi phục vụ phân tích (ví dụ tài khoản gửi, tài khoản nhận hoặc giá trị giao dịch) được loại bỏ khỏi tập dữ liệu nhằm bảo đảm tính nhất quán của các đặc trưng hành vi.

Đối với các thuộc tính số còn lại, giá trị thiếu được thay thế bằng **giá trị trung vị (median)** của thuộc tính tương ứng, thay vì giá trị trung bình, nhằm giảm ảnh hưởng của các phân bố lệch và ngoại lai, vốn là đặc điểm phổ biến trong dữ liệu tài chính. Cách xử lý này góp phần duy trì hình dạng phân bố ổn định của dữ liệu và tránh làm sai lệch các đặc trưng hành vi quan trọng.

2.2.3.2 Chuẩn hóa thang đo và xử lý phân bố lệch

Dữ liệu giao dịch tài chính thường có phân bố không đối xứng và xuất hiện các giá trị ngoại lai lớn, đặc biệt đối với các thuộc tính như số tiền giao dịch hoặc thay đổi số dư. Nếu đưa trực tiếp các biến này vào mô hình, các giá trị cực đoan có thể chi phối quá trình học và làm suy giảm khả năng tổng quát hóa.

Để khắc phục vấn đề này, các bước xử lý sau được áp dụng:

Biến đổi log cho các thuộc tính có phân bố lệch mạnh:

Với các biến dương có đuôi dài (ví dụ `amount`), phép biến đổi log được áp dụng:

$$x' = log(1 + x)$$

Phép biến đổi này giúp:

- Giảm độ lệch của phân bố
- Nén miền giá trị
- Giữ nguyên thứ tự tương đối giữa các giao dịch

Chuẩn hóa thang đo:

Sau khi biến đổi log (nếu cần), các thuộc tính số được chuẩn hóa về cùng thang đo nhằm hỗ trợ quá trình huấn luyện mô hình. Việc chuẩn hóa giúp tránh tình trạng một số đặc trưng chi phối hoàn toàn gradient trong các mô hình học sâu, đặc biệt là khi kết hợp nhiều nguồn đặc trưng khác nhau.

Trong đề án, chuẩn hóa được thực hiện theo hướng **bền vững với ngoại lai**, dựa trên trung vị và khoảng tứ phân vị (Interquartile Range – IQR):

$$x_{scaled} = \frac{x - median(x)}{IQR(x)}$$

Việc chuẩn hóa chỉ được **khớp trên tập huấn luyện**, sau đó áp dụng nguyên vẹn cho tập kiểm định và tập kiểm tra nhằm tránh rò rỉ thông tin.

2.2.3.3 Xây dựng đặc trưng thời gian (Temporal Encoding)

Thời điểm phát sinh giao dịch là một nguồn thông tin quan trọng trong bài toán phát hiện gian lận, do hành vi gian lận thường gắn với các chu kỳ hoặc khung thời gian đặc thù (ví dụ ban đêm, cuối tuần, hoặc các giai đoạn biến động).

Trong đề án, thời gian được mã hóa thông qua tập đặc trưng sau:

Chuẩn hóa thời gian tuyến tính

Thời gian giao dịch ttt được chuẩn hóa về khoảng $[0, 1]$ theo công thức:

$$t_{norm} = \frac{t - t_{min}}{t_{max} - t_{min}}$$

trong đó t là thời điểm giao dịch, t_{min} và t_{max} lần lượt là thời điểm sớm nhất và muộn nhất trong tập dữ liệu.

Biểu diễn này giúp mô hình học được vị trí tương đối của mỗi giao dịch trong toàn bộ dòng thời gian quan sát.

Mã hóa chu kỳ (cyclic encoding)

Để biểu diễn tính chu kỳ của thời gian, hai đặc trưng được xây dựng:

$$t_{sin} = sin(2\pi \cdot t_{norm}), \; t_{cos} = cos(2\pi \cdot t_{norm})$$

Cách mã hóa này cho phép mô hình học được sự liên tục giữa các thời điểm liền kề trong chu kỳ (ví dụ cuối ngày và đầu ngày), điều mà biểu diễn tuyến tính thuần túy không thể hiện hiệu quả.

Phân nhóm thời gian (time bucket)

Ngoài ra, thời gian còn được ánh xạ vào các khoảng rời rạc thông qua biến t_{bucket} nhằm phản ánh các pha hoạt động khác nhau trong ngày hoặc trong chu kỳ giao dịch.

$$t_{bucket} = \left[\frac{t - t_{min}}{\Delta t}\right]$$

trong đó Δt là độ dài mỗi khoảng thời gian (ví dụ theo giờ hoặc theo ngày). Đặc trưng này giúp mô hình phân biệt các pha hoạt động khác nhau trong ngày hoặc trong chu kỳ giao dịch.

2.2.3.4 Phân chia dữ liệu theo thứ tự thời gian

Sau khi hoàn tất các bước làm sạch và xây dựng đặc trưng, dữ liệu được phân chia theo thứ tự thời gian thành ba tập:

- **Tập huấn luyện (Train):** 70% dữ liệu sớm nhất
- **Tập kiểm định (Validation):** 15% dữ liệu tiếp theo
- **Tập kiểm tra (Test):** 15% dữ liệu muộn nhất

Việc phân chia theo thời gian nhằm mô phỏng đúng kịch bản triển khai thực tế, trong đó mô hình chỉ có thể học từ dữ liệu quá khứ để dự đoán các giao dịch phát sinh trong tương lai. Cách tiếp cận này giúp tránh rò rỉ thông tin và bảo đảm rằng

các chỉ số đánh giá phản ánh đúng khả năng tổng quát hóa của pipeline.

2.2.3.5 Vai trò của tiền xử lý trong cấu trúc tổng thể

Quy trình tiền xử lý dữ liệu không chỉ đóng vai trò chuẩn bị dữ liệu đầu vào cho mô hình, mà còn quyết định trực tiếp đến hiệu quả của các bước tiếp theo trong pipeline. Cụ thể:

- Các đặc trưng đã được chuẩn hóa và mã hóa thời gian là đầu vào cho bước **gom cụm tài khoản (account clustering)**, giúp việc phân nhóm dựa trên hành vi trở nên ổn định hơn.
- Dữ liệu sau tiền xử lý được sử dụng để xây dựng đồ thị giao dịch, trong đó các node và edge mang các đặc trưng nhất quán về thang đo và phân bố.
- Việc tách dữ liệu theo thời gian bảo đảm rằng các bước học biểu diễn bằng GNN và xử lý mất cân bằng dữ liệu bằng SMOTE được đánh giá trong bối cảnh sát với vận hành thực tế.

Tổng thể, bước tiền xử lý được xem là nền tảng kỹ thuật cho toàn bộ pipeline phát hiện gian lận trong đề án, bảo đảm tính nhất quán, khả năng tái lập và độ tin cậy của các kết quả thực nghiệm.

2.2.4 Tóm tắt mô hình tiền xử lý dữ liệu

Sau khi mô tả chi tiết từng bước tiền xử lý ở mục 2.2.3, phần này tổng hợp lại toàn bộ quy trình dưới dạng một cấu trúc thống nhất, nhằm làm rõ luồng dữ liệu từ đầu vào thô đến đầu ra sẵn sàng cho các bước xây dựng đồ thị, học biểu diễn và phân loại.

Bảng 2.2: Quy trình tiền xử lý dữ liệu



Cách tổ chức mô hình theo chuỗi giúp bảo đảm rằng mọi phép biến đổi đều được thực hiện có kiểm soát, đồng thời dễ dàng tái lập trong các thí nghiệm mở rộng hoặc triển khai

thực tế.

2.2.5 Định dạng dữ liệu sau tiền xử lý

Sau khi hoàn tất phân đoạn tiền xử lý, dữ liệu được xuất ra các định dạng cụ thể để phục vụ các bước tiếp theo trong đề án.

Định dạng dữ liệu bảng (Tabular format)

Dữ liệu giao dịch sau tiền xử lý được lưu dưới dạng bảng, trong đó mỗi dòng tương ứng với một giao dịch và mỗi cột tương ứng với một đặc trưng đã chuẩn hóa. Định dạng này được sử dụng cho:

- Gom cụm tài khoản (account clustering) dựa trên hành vi giao dịch
- Xây dựng đặc trưng node và edge làm cơ sở tạo đồ thị giao dịch.
- Huấn luyện bộ phân loại Random Forest, đặc biệt trong nhánh thực nghiệm có áp dụng SMOTE trên embedding giao dịch.

Định dạng dữ liệu đồ thị (Graph-ready format)

Từ dữ liệu bảng đã tiền xử lý, các cặp tài khoản gửi–nhận được ánh xạ sang dạng:

- Danh sách node (cụm tài khoản)
- Danh sách edge (giao dịch giữa các cụm)

Mỗi node và edge đều gắn kèm vector đặc trưng đã chuẩn hóa, bảo đảm tính nhất quán khi đưa vào mô hình GNN.

Mục đích:

Việc tách rõ định dạng dữ liệu đầu ra giúp pipeline có tính mô-đun, cho phép thay thế hoặc mở rộng từng thành phần (ví dụ thay đổi mô hình GNN hoặc bộ phân loại) mà không cần thiết kế lại toàn bộ quy trình tiền xử lý.

Trên cơ sở dữ liệu đã được tiền xử lý và chuẩn hóa theo pipeline trình bày ở các mục trên, bước tiếp theo của đề án là tổ chức lại các giao dịch dưới dạng **cấu trúc đồ thị** nhằm khai thác mối quan hệ giữa các thực thể giao dịch.

Do đặc thù dữ liệu giao dịch có số lượng tài khoản lớn và mức độ kết nối không đồng đều, đề án **không xây dựng đồ thị trực tiếp trên từng tài khoản riêng lẻ**, mà áp dụng bước **gom cụm tài khoản (account clustering)** như một cơ chế trung gian nhằm:

- Giảm độ rời rạc của đồ thị.
- Tăng mức độ kết nối giữa các node.
- Tạo điều kiện thuận lợi cho việc học biểu diễn quan hệ bằng GNN.

Phần tiếp theo của chương sẽ trình bày chi tiết phương pháp gom cụm tài khoản, tiêu chí lựa chọn số cụm và cách xây dựng đồ thị giao dịch dựa trên các cụm tài khoản này.

2.3 Biểu diễn dữ liệu và xây dựng đồ thị

2.3.1 Gom cụm tài khoản (Account Clustering)

Trước khi thực hiện gom cụm, dữ liệu giao dịch sau tiền xử lý được **tổng hợp ở mức tài khoản** nhằm xây dựng các vector đặc trưng hành vi làm đầu vào cho thuật toán K-Means. Cụ thể, mỗi tài khoản được biểu diễn bởi một vector đặc trưng tổng hợp từ toàn bộ các giao dịch liên quan trong tập huấn luyện, bao gồm các thống kê về tần suất giao dịch, tổng và trung bình giá trị giao dịch, cũng như các đặc trưng thời gian đã mã hóa. Bước lập dữ liệu này cho phép chuyển đổi dữ liệu giao dịch rời rạc thành biểu diễn hành vi ổn định ở mức tài khoản, làm cơ sở cho việc gom cụm và giảm độ nhiễu do các giao dịch đơn lẻ gây ra.

Sau đó mới gom cụm tài khoản (account clustering) - là bước ánh xạ tập tài khoản gốc về một tập hữu hạn các nhóm hành vi, trong đó mỗi cụm đại diện cho một nhóm tài khoản có đặc trưng giao dịch tương đồng. Đầu vào của bước này là các vector đặc trưng hành vi ở mức tài khoản (được tổng hợp từ dữ liệu giao dịch sau tiền xử lý như đã nói ở trên); đầu ra là nhãn cụm cho từng tài khoản và tập các cụm dùng để xây dựng node của đồ thị. Trên cơ sở định nghĩa này, phần tiếp theo trình bày động cơ sử dụng Account Clustering, chi tiết thuật toán gom cụm và tiêu

chí lựa chọn số cụm.

Động cơ sử dụng gom cụm tài khoản

Một thách thức lớn khi áp dụng trực tiếp các mô hình GNN lên dữ liệu giao dịch PaySim là cấu trúc đồ thị ở mức tài khoản nguyên bản rất rời rạc. Tập dữ liệu chứa số lượng tài khoản lớn, trong đó phần đáng kể các tài khoản chỉ xuất hiện với tần suất rất thấp (một hoặc vài giao dịch). Khi biểu diễn trực tiếp mỗi tài khoản như một node và mỗi giao dịch như một edge, đồ thị thu được có mật độ kết nối thấp, phân mảnh mạnh và thiếu các cấu trúc lân cận đủ dày để cơ chế truyền thông tin (message passing) của GNN phát huy hiệu quả.

Trong bối cảnh này, việc lan truyền thông tin qua nhiều tầng GNN trở nên kém hiệu quả do mỗi node chỉ có rất ít hàng xóm, dẫn đến embedding học được mang tính cục bộ và khó phản ánh các mẫu hành vi gian lận mang tính liên kết hoặc theo chuỗi. Đồng thời, số lượng node quá lớn cũng làm tăng chi phí tính toán và gây khó khăn trong quá trình huấn luyện mô hình.

Để khắc phục những hạn chế trên, đề án áp dụng bước gom cụm tài khoản (account clustering) như một cơ chế trung gian trước khi xây dựng đồ thị. Thay vì coi mỗi tài khoản là một thực thể độc lập, các tài khoản có hành vi giao dịch tương đồng được gom lại thành các cụm dựa trên các đặc trưng hành vi tổng hợp đã được tiền xử lý và chuẩn hóa.

Cách tiếp cận này mang lại các lợi ích chính sau:

- Giảm số lượng node trong đồ thị xuống mức phù hợp với khả năng học biểu diễn của GNN, từ đó cải thiện hiệu quả tính toán và khả năng hội tụ của mô hình.
- Tăng mật độ kết nối giữa các node, do các giao dịch giữa nhiều tài khoản khác nhau nay được quy tụ thành các tương tác giữa các cụm, giúp cơ chế truyền thông tin của GNN hoạt động hiệu quả hơn.

- Xây dựng một biểu diễn mang tính xấp xỉ thực tế, trong đó mỗi cụm tài khoản đại diện cho một nhóm hành vi giao dịch tương đồng, phản ánh sát hơn các mô thức gian lận theo nhóm hoặc theo vai trò chức năng thường gặp trong thực tế.

Trong thực nghiệm của đề án, thuật toán K-Means được sử dụng để gom cụm tài khoản với số cụm cố định. Việc lựa chọn số cụm nhằm đạt được sự cân bằng giữa hai yếu tố: (i) giữ đủ độ chi tiết để phân biệt các nhóm hành vi khác nhau, và (ii) duy trì cấu trúc đồ thị đủ dày để GNN có thể học được các quan hệ có ý nghĩa.

2.3.2 Mô hình gom cụm tài khoản bằng K-Means

K-Means là thuật toán phân cụm dựa trên khoảng cách, nhằm chia tập các điểm dữ liệu thành K cụm sao cho tổng sai số nội cụm được tối thiểu hóa. Thuật toán vận hành theo cơ chế lặp giữa (i) gán điểm về tâm gần nhất và (ii) cập nhật tâm cụm bằng trung bình các điểm trong cụm.

Trong ngữ cảnh pipeline, mỗi "điểm dữ liệu" tương ứng với một tài khoản và được biểu diễn bởi vector đặc trưng hành vi sau tiền xử lý $x_i \in R^d$, trong đó d là số lượng đặc trưng mô tả hành vi giao dịch của tài khoản sau tiền xử lý (ví dụ: thống kê giá trị giao dịch, tần suất, biến thời gian đã mã hóa) ; "cụm" tương ứng với nhóm tài khoản có hành vi tương đồng và sẽ được dùng làm node của đồ thị giao dịch.

Bài toán **K-Means** được phát biểu như sau:

Cho tập N tài khoản với các vector đặc trưng $\{x_1, x_2, ..., x_N\}$, mục tiêu là tìm tập K tâm cụm

$$\{\mu_1, \mu_2, ..., \mu_N\}$$

sao cho hàm mất mát nội cụm được tối thiểu hóa:

$$\iota = \sum_{k=1}^{K} \sum_{x_i \in C_k} \left|\left| x_i - \mu_k \right|\right|_2^2$$

Trong đó:

- C_k là tập các tài khoản thuộc cụm thứ k,
- μ_k là vector trung tâm (centroid) của cụm k,
- $\| \cdot \|_2$ là chuẩn Euclid.

Thuật toán **K-Means** lặp lại hai bước cho đến khi hội tụ:

1. **Gán cụm**: mỗi tài khoản x_i được gán vào cụm có tâm gần nhất:

$$cluster(x_i) = arg\,min_k \left|\left| x_i - \mu_k \right|\right|_2$$

2. **Cập nhật tâm cụm:**

$$\mu_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i$$

Lựa chọn số cụm $K = 200$

Trong thực nghiệm của đề án, số cụm được **cố định là** $K = 200$

Việc lựa chọn giá trị này dựa trên các cân nhắc sau:

Đầu tiên, dữ liệu không được sử dụng trực tiếp dưới một dạng duy nhất mà được **tổ chức thành các tập dữ liệu trung gian tương ứng với từng giai đoạn học**. Cụ thể, (i) tập dữ liệu hành vi tài khoản được sử dụng cho bước gom cụm K-Means, (ii) tập dữ liệu đồ thị với node là các cụm tài khoản và edge là các giao dịch giữa cụm được sử dụng cho mô hình GNN, và (iii) tập dữ liệu embedding giao dịch (edge embedding) được sử dụng cho bài toán phân loại gian lận. Cách tổ chức này giúp tách bạch rõ vai trò của từng giai đoạn, đồng thời tăng tính mộ-đun và khả năng tái lập của pipeline.

Dựa trên quy mô dữ liệu PaySim (hơn 6,3 triệu giao dịch và số lượng tài khoản lớn), cũng như tham khảo các thiết lập trong các nghiên cứu khai thác hành

vi tài khoản ở mức nhóm, phạm vi giá trị K được khoanh vùng trong khoảng **[50, 400]** cho quá trình thử nghiệm.

Kết quả thử nghiệm cho thấy:

- Khi K **nhỏ (ví dụ** $K < 100$**)**, các cụm trở nên quá thô, nhiều tài khoản có hành vi khác biệt bị gom chung vào cùng một cụm. Điều này làm suy giảm khả năng phân biệt hành vi và dẫn đến giảm hiệu quả phát hiện gian lận, thể hiện qua việc PR-AUC và F1-score thấp hơn.
- Khi K **quá lớn (ví dụ** $K > 300$**)**, số lượng node trong đồ thị tăng lên đáng kể, khiến đồ thị trở nên thưa và phân mảnh. Trong trường hợp này, hiệu quả lan truyền thông tin của GNN giảm, đồng thời chi phí tính toán tăng, trong khi hiệu năng mô hình không cải thiện tương ứng và có xu hướng dao động.
- Khi K **nằm trong khoảng trung bình (xung quanh 200)**:

  Giá trị $K = 200$ cho phép đạt được sự cân bằng giữa:

  - **Độ chi tiết hành vi**: các cụm vẫn đủ nhỏ để phản ánh các nhóm hành vi khác nhau;
  - **Mật độ kết nối đồ thị**: số node giảm đáng kể so với số tài khoản ban đầu, giúp đồ thị đủ dày để GNN học được quan hệ.
  - **Hiệu quả phát hiện gian lận** của pipeline thông qua các chỉ số PR-AUC và F1-score.

2.3.3 Xây dựng đồ thị giao dịch theo cụm

Đồ thị giao dịch trong pipeline đề xuất là một biểu diễn có hướng của luồng chuyển tiền, trong đó node không còn là tài khoản đơn lẻ mà là các cụm tài khoản sau bước gom cụm. Đầu vào của bước này gồm: nhãn cụm của tài khoản và danh sách giao dịch gửi–nhận; đầu ra là đồ thị $G = (V, E)$ với tập node V là các cụm và tập cạnh E là các luồng giao dịch giữa các cụm, kèm theo vector đặc trưng cho node và edge. Việc biểu diễn dữ liệu dưới dạng đồ thị cho phép khai thác cấu trúc quan

hệ và các mẫu liên kết mà dữ liệu bảng khó phản ánh đầy đủ.

Trên cơ sở đó, sau khi hoàn tất bước gom cụm tài khoản, dữ liệu giao dịch được tổ chức lại dưới dạng một đồ thị có hướng:

$$\varsigma = (\gamma, \varepsilon)$$

Trong đó:

- Tập node $\gamma = \{1, 2, ..., K\}$ tương ứng với các **cụm tài khoản**;
- Mỗi edge có hướng

$$e = (u \to \upsilon) \in \varepsilon$$

tương ứng với một giao dịch từ tài khoản thuộc cụm u sang tài khoản thuộc cụm υ.

Mỗi node u được gán một vector đặc trưng:

$$e_{u\upsilon} \in R^{d_e}$$

mô tả hành vi tổng hợp của cụm (thống kê giao dịch, đặc trưng thời gian, …).

Mỗi edge $(u \to \upsilon)$ mang vector đặc trưng:

$$h_u^{(0)} \in R^{d_n}$$

mô tả thông tin của giao dịch như giá trị, loại giao dịch và thời điểm phát sinh (sau tiền xử lý).

Biểu diễn này cho phép mô hình hóa trực tiếp **luồng giao dịch giữa các nhóm tài khoản**, thay vì xử lý từng giao dịch như các mẫu độc lập, từ đó tạo nền tảng cho việc học biểu diễn bằng GNN ở các bước tiếp theo.

2.3.4 Vai trò của biểu diễn đồ thị trong mô hình tổng thể

Biểu diễn đồ thị theo cụm tài khoản đóng vai trò trung gian giữa bước tiền xử lý và bước học biểu diễn sâu. Nhờ cấu trúc đồ thị này, các mô hình GNN trong đề án có thể khai thác đồng thời:

- Đặc trưng hành vi của từng cụm tài khoản,

- Thông tin chi tiết của các giao dịch,
- Cấu trúc liên kết toàn cục của mạng lưới giao dịch.

Trên nền tảng đó, bước tiếp theo của pipeline tập trung vào việc học embedding bằng **Graph Attention Network**, cho phép gán trọng số khác nhau cho các mối quan hệ giao dịch và làm nổi bật các tương tác mang tính bất thường.

Để bảo đảm tính công bằng trong đánh giá và tránh rò rỉ thông tin theo thời gian, toàn bộ các bước lập dữ liệu mang tính học tham số, bao gồm gom cụm tài khoản và học biểu diễn đồ thị, đều được thực hiện **dựa trên tập huấn luyện**. Các nhãn cụm và cấu trúc biểu diễn thu được sau đó được áp dụng nhất quán cho tập kiểm định và tập kiểm tra, trong khi các tập này chỉ được sử dụng cho mục đích đánh giá. Nguyên tắc này bảo đảm rằng các kết quả so sánh giữa các mô hình phản ánh đúng khả năng tổng quát hóa của mô hình trong kịch bản triển khai thực tế.

2.4 Học biểu diễn và mô hình phân loại

Sau khi hoàn tất các bước tiền xử lý và xây dựng cấu trúc đồ thị giao dịch theo cụm tài khoản, bước tiếp theo của pipeline là học biểu diễn (representation learning) cho các thực thể trong đồ thị và xây dựng mô hình phân loại phục vụ phát hiện gian lận. Mục tiêu của giai đoạn này là khai thác đồng thời thông tin đặc trưng của từng cụm tài khoản và thông tin quan hệ giữa các cụm, từ đó tạo ra các biểu diễn giàu ngữ nghĩa cho giao dịch và hỗ trợ phân loại hiệu quả trong bối cảnh dữ liệu cực kỳ mất cân bằng.

2.4.1 Học embedding bằng NodeMLP

NodeMLP ở đây là một mạng nơ-ron đa tầng dùng để ánh xạ đặc trưng hành vi ban đầu của mỗi node (cụm tài khoản) từ không gian dữ liệu bảng sang một không gian embedding liên tục có chiều cố định. Đầu vào của NodeMLP là vector đặc trưng node x_u sau tiền xử lý/tổng hợp; đầu ra là embedding h_u đóng vai trò biểu diễn khởi tạo cho GNN. Bước này giúp mô hình hóa các quan hệ phi tuyến giữa các đặc trưng hành vi và chuẩn hóa dạng biểu diễn trước khi lan truyền thông tin theo

cấu trúc đồ thị.

Trên cơ sở lý thuyết này, sau khi xây dựng đồ thị giao dịch dựa trên các cụm tài khoản, mỗi node (cụm tài khoản) được gán một vector đặc trưng ban đầu tổng hợp từ hành vi giao dịch. Để chuyển các đặc trưng dạng bảng này sang không gian biểu diễn liên tục phù hợp cho mô hình đồ thị, đề án sử dụng một mạng nơ-ron đa tầng (NodeMLP).

Gọi:

- $x_u \in R^{d_0}$ là vector đặc trưng ban đầu của node(cụm tài khoản) u
- L là số tầng của mạng MLP

Biểu diễn của node sau NodeMLP được tính theo dạng:

- Tầng 1:

$$h_u^{(1)} = \sigma\left(W^{(1)} x_u + b^{(1)}\right)$$

- Tầng $l - th$ (với $l = 2, ..., L$)

$$h_u^{(l)} = \sigma\left(W^{(l)} h_u^{(l-1)} + b^{(l)}\right)$$

Trong đó:

- $W^{(l)}$, $b^{(l)}$ là trọng số và bias tại tầng l,
- $\sigma(.)$ là hàm kích hoạt phi tuyến (ReLU),
- $h_u^{(l)} \in R^d$ là embedding cuối cùng của node u.

NodeMLP đóng vai trò như một tầng học biểu diễn cục bộ, cho phép mô hình tự động trích xuất các tổ hợp phi tuyến giữa các đặc trưng hành vi của cụm tài khoản trước khi đưa vào GNN.

2.4.2 Mô hình hóa quan hệ bằng Graph Attention Network

GAT là mô hình học trên đồ thị dùng để cập nhật embedding node bằng

cách tổng hợp thông tin từ các node lân cận với trọng số khác nhau, trong đó trọng số được học tự động thông qua cơ chế attention. Đầu vào của GAT là đồ thị giao dịch $G = (V, E)$ và embedding node khởi tạo (từ NodeMLP); đầu ra là embedding node đã tích hợp thông tin quan hệ, phản ánh đồng thời đặc trưng nội tại và cấu trúc liên kết. Việc sử dụng attention cho phép mô hình phân biệt mức độ quan trọng của các luồng giao dịch khác nhau, phù hợp với đặc thù dữ liệu tài chính nơi không phải mọi liên kết đều mang cùng mức độ rủi ro.

Nói ngắn gọn, sau khi mỗi node được mã hóa thành embedding ban đầu thông qua NodeMLP, GAT được sử dụng để học biểu diễn quan hệ giữa các node trong đồ thị giao dịch.

Gọi:

- $N(u)$ là tập các node lân cận của node u
- $h_u^{(l)}$ là embedding của node u tại tầng GAT thứ l.

2.4.2.1 Cơ chế attention

Với mỗi cạnh (u, v) hệ số attention chưa chuẩn hóa được tính như sau:

$$e_{uv} = LeakyReLU\left(a^T\left[Wh_u || Wh_u\right]\right)$$

Trong đó:

- W là ma trận biến đổi tuyến tính,
- a là vector tham số attention,
- $||$ là phép nối vector.

Hệ số attention được chuẩn hóa bằng softmax:

$$\alpha_{uv} = \frac{exp\left(e_{uv}\right)}{\sum_{k \in N(u)} exp\left(e_{uk}\right)}$$

2.4.2.2 Cập nhật embedding node

Embedding mới của node u được tính bằng tổng có trọng số các node lân cận:

$$h_u^{'} = \sigma\left(\sum_{\upsilon \in N(u)} \alpha_{u\upsilon} W h_\upsilon\right)$$

Với multi-head attention (H head), biểu diễn cuối cùng là phép nối hoặc trung bình các head:

$$h_u^{(l+1)} = ||_{i=1}^{H} h_{u,i}^{'}$$

Cơ chế attention cho phép mô hình tập trung vào các mối quan hệ giao dịch quan trọng, phản ánh thực tế rằng chỉ một số liên kết đóng vai trò then chốt trong hành vi gian lận.

2.4.3 Edge embedding và biểu diễn giao dịch

Sau khi thu được biểu diễn nhúng cuối cùng của các node từ GAT, mỗi giao dịch (cạnh) giữa hai cụm tài khoản được biểu diễn thông qua vector biểu diễn của node nguồn và node đích, gọi là edge embedding- biểu diễn nhúng của giao dịch.

Gọi:

- h_u là embedding của node nguồn,
- h_υ là embedding của node đích.

Embedding của giao dịch $(u \to \upsilon)$ được xây dựng như sau:

$$e_{u,\upsilon} = \left[h_u \,||\, h_\upsilon \,||\, \left|h_u - h_\upsilon\right| \,||\, \left(h_u \odot h_\upsilon\right)\right]$$

Trong đó:

- $\left|h_u - h_\upsilon\right|$ là độ chênh lệch tuyệt đối,
- $\odot$ là phép nhân theo phần tử.

Cách kết hợp này cho phép embedding giao dịch nắm bắt:

- thông tin từng phía (nguồn và đích),
- mức độ khác biệt hành vi giữa hai cụm,
- mức độ tương tác giữa các đặc trưng.

2.4.4 Xử lý mất cân bằng dữ liệu bằng SMOTE

Dữ liệu giao dịch tài chính có đặc điểm nổi bật là tỷ lệ gian lận cực thấp. Nếu huấn luyện mô hình phân loại trực tiếp trên tập dữ liệu gốc, mô hình có xu hướng thiên lệch mạnh về lớp không gian lận và khó học được ranh giới phân lớp hiệu quả cho lớp thiểu số.

Để đánh giá tác động của kỹ thuật cân bằng dữ liệu, đề án triển khai hai nhánh thực nghiệm song song:

- **Nhánh baseline:** mô hình được huấn luyện trực tiếp trên embedding giao dịch gốc, không áp dụng bất kỳ kỹ thuật cân bằng dữ liệu nào.
- **Nhánh chính:** kỹ thuật SMOTE được áp dụng trên không gian embedding của giao dịch trong tập huấn luyện nhằm mở rộng lớp gian lận.

Với nhánh chính áp dụng SMOTE:

SMOTE (Synthetic Minority Over-sampling Technique) trong pipeline đề xuất là kỹ thuật tăng mẫu lớp thiểu số bằng cách sinh các điểm tổng hợp thông qua nội suy giữa các mẫu dương tính gần nhau trong không gian đặc trưng. Khác với cách áp dụng SMOTE trên đặc trưng thô, luận văn áp dụng SMOTE trên **embedding giao dịch** (edge embedding) đã tích hợp thông tin quan hệ từ GNN. Đầu vào của bước này là tập embedding giao dịch thuộc lớp gian lận trong tập huấn luyện; đầu ra là các embedding tổng hợp bổ sung cho lớp thiểu số, giúp bộ phân loại học ranh giới phân lớp tốt hơn trong bối cảnh tỷ lệ gian lận cực thấp.

Gọi:

- x_i là một embedding giao dịch thuộc lớp gian lận,
- x_{nn} là một trong các láng giềng gần nhất của x_i trong không gian

embedding.

Một mẫu tổng hợp mới được sinh theo công thức:

$$x_{new} = x_i + \lambda\left(x_{nn} - x_i\right),\ \lambda \sim U(0,\ 1)$$

Trong đó: λ là hệ số ngẫu nhiên phân bố đều trong khoảng $[0,\ 1]$.

Trong pipeline đề xuất, SMOTE chỉ được áp dụng **sau khi hoàn tất bước học embedding giao dịch** và **chỉ trên tập huấn luyện**, trong khi các tập kiểm định và kiểm tra được giữ nguyên phân bố dữ liệu ban đầu. Thiết kế này bảo đảm rằng việc cân bằng dữ liệu không gây rò rỉ thông tin và cho phép so sánh công bằng giữa nhánh baseline và nhánh có áp dụng SMOTE. Việc áp dụng SMOTE trên embedding (thay vì đặc trưng thô) giúp các mẫu tổng hợp nằm trong không gian đã phản ánh ngữ cảnh quan hệ của đồ thị, giảm nguy cơ tạo ra các mẫu gian lận không thực tế.

2.4.5 Bộ phân loại Random Forest

Embedding giao dịch sau khi cân bằng (hoặc không cân bằng, trong nhánh baseline) được đưa vào bộ phân loại RF.

RF là bộ phân loại ensemble gồm nhiều cây quyết định, trong đó mỗi cây được huấn luyện trên một mẫu bootstrap của dữ liệu và một tập con đặc trưng ngẫu nhiên. Đầu vào của RF là embedding giao dịch (có hoặc không qua SMOTE tùy nhánh thực nghiệm); đầu ra là xác suất/nhãn gian lận cho từng giao dịch.

Trong pipeline, RF gồm T cây quyết định, trong đó mỗi cây $f_t(\cdot)$ được huấn luyện trên một mẫu bootstrap của dữ liệu.

Dự đoán xác suất gian lận cho một giao dịch được tính bằng trung bình các cây:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t\left(e_{uv}\right)$$

Trong đó: $f_t(e_{uv}) \in \{0, 1\}$ hoặc là xác suất lớp gian lận tại cây thứ t.

RF cho phép học các ranh giới phân lớp phi tuyến trong không gian embedding nhiều chiều, đồng thời có khả năng chống nhiễu tốt và ổn định với dữ liệu mất cân bằng.

2.5 Thiết kế huấn luyện và phân tích

2.5.1 Pipeline huấn luyện

Pipeline huấn luyện được thiết kế thống nhất cho cả hai nhánh thực nghiệm (có và không áp dụng SMOTE). Quy trình bao gồm các bước: nạp dữ liệu đã tiền xử lý, học embedding node bằng NodeMLP, lan truyền thông tin bằng GAT, xây dựng edge embedding, xử lý mất cân bằng (nếu có), huấn luyện bộ phân loại RF và đánh giá mô hình.

Trong đó, kỹ thuật SMOTE chỉ được áp dụng trên tập huấn luyện nhằm mở rộng lớp gian lận, trong khi các tập kiểm định và kiểm tra được giữ nguyên phân bố dữ liệu ban đầu để phản ánh đúng điều kiện vận hành thực tế. Các tham số của NodeMLP, GAT và RF được cố định sau giai đoạn huấn luyện và không điều chỉnh trong quá trình đánh giá trên tập kiểm tra.

Chuẩn hóa pipeline huấn luyện giúp bảo đảm tính công bằng khi so sánh hiệu quả giữa các thiết lập, cũng cho phép tái lập kết quả trong các thí nghiệm mở rộng.

2.5.2 Các thước đo đánh giá

Do dữ liệu gian lận có tỷ lệ dương tính rất thấp, các thước đo dựa trên xếp hạng và đánh đổi giữa precision và recall được ưu tiên. Các chỉ số chính bao gồm:

- ROC-AUC: đánh giá khả năng phân biệt tổng quát giữa hai lớp.
- PR-AUC: phản ánh trực tiếp hiệu quả phát hiện lớp gian lận hiếm.
- F1-score: đánh giá cân bằng giữa precision - recall tại ngưỡng phân loại.
- Precision và Recall: hỗ trợ phân tích chi tiết hành vi mô hình.

Việc đánh giá được thực hiện trên tập kiểm tra chưa từng được mô hình nhìn thấy, với toàn bộ mô hình và ngưỡng phân loại được xác định trước nhằm phản ánh đúng khả năng tổng quát hóa của pipeline trong kịch bản triển khai thực tế.

2.6 Kết luận chương

Chương hai đã trình bày chi tiết phương pháp nghiên cứu và thiết kế thực nghiệm của đề án, từ bước học biểu diễn dựa trên đồ thị đến mô hình phân loại và chiến lược đánh giá. Các quyết định thiết kế đều bám sát đặc thù của dữ liệu giao dịch tài chính, đặc biệt là tính liên kết giữa các thực thể và mức độ mất cân bằng nghiêm trọng của nhãn gian lận.

CHƯƠNG 3: TRIỂN KHAI THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

3.1 Giới thiệu chung

Say đây là quá trình triển khai thực nghiệm nhằm đánh giá hiệu quả của pipeline phát hiện gian lận giao dịch tài chính được đề xuất ở Chương 2. Trọng tâm của chương là kiểm chứng vai trò của biểu diễn đồ thị và kỹ thuật xử lý mất cân bằng dữ liệu trong bối cảnh dữ liệu giao dịch có tỷ lệ gian lận cực thấp.

Các thí nghiệm được thiết kế xoay quanh **hai pipeline** có cùng cấu trúc học biểu diễn nhưng khác nhau ở bước xử lý mất cân bằng dữ liệu, bao gồm:

- **Mô hình baseline**: NodeMLP → GAT →RF, huấn luyện trực tiếp trên dữ liệu gốc không cân bằng.
- **Mô hình chính**: NodeMLP → GAT → RF kết hợp với **SMOTE** trên không gian embedding của giao dịch.

Thông qua việc so sánh hai mô hình này, chương 3 muốn làm rõ:

- Sự khác biệt giữa các thước đo tổng quát (ROC-AUC) và các thước đo tập trung vào lớp gian lận hiếm (PR-AUC, F1-score).
- Tác động thực tế của kỹ thuật cân bằng dữ liệu khi kết hợp với biểu diễn đồ thị trong bài toán phát hiện gian lận.

3.2 Môi trường và công cụ thực nghiệm

Toàn bộ quá trình tiền xử lý dữ liệu, xây dựng đồ thị, huấn luyện và đánh giá mô hình trong đề án được triển khai trên nền tảng Google Colaboratory (Google Colab). Việc sử dụng Colab cho phép chuẩn hóa môi trường thực nghiệm, giảm phụ thuộc vào cấu hình phần cứng cục bộ và hỗ trợ khả năng tái lập kết quả.

Các thí nghiệm được thực hiện trong môi trường Python, tận dụng hệ sinh thái thư viện mã nguồn mở phục vụ học máy và học sâu. Đối với các bước huấn luyện mô hình học sâu như NodeMLP và GAT, đề án sử dụng GPU do Colab cung cấp nhằm

tăng tốc tính toán và rút ngắn thời gian huấn luyện.

Tập dữ liệu PaySim được cung cấp dưới dạng các tệp dữ liệu bảng (CSV), được lưu trữ trên Google Drive và nạp vào môi trường thực nghiệm thông qua cơ chế mount của Google Colab, sử dụng thư viện Pandas để đọc và xử lý dữ liệu.

Các thư viện chính được sử dụng bao gồm:

- NumPy, Pandas cho xử lý và phân tích dữ liệu,
- Scikit-learn cho các bước tiền xử lý và mô hình học máy truyền thống,
- PyTorch/PyTorch Geometric cho xây dựng và huấn luyện mô hình đồ thị,
- Imbalanced-learn cho kỹ thuật cân bằng dữ liệu SMOTE,
- Matplotlib cho trực quan hóa kết quả thực nghiệm.

Việc triển khai trên Google Colab phản ánh đúng bối cảnh nghiên cứu học thuật, trong đó trọng tâm đặt vào đánh giá phương pháp và phân tích mô hình, thay vì tối ưu hạ tầng triển khai.

3.3 Thiết lập thực nghiệm

3.3.1 Dữ liệu và phân chia tập

Thực nghiệm được tiến hành trên tập dữ liệu giao dịch tài chính **PaySim**, là dữ liệu mô phỏng dựa trên nhật ký giao dịch của hệ thống thanh toán di động. Tập dữ liệu bao gồm hơn 6,3 triệu giao dịch, trong đó tỷ lệ giao dịch gian lận nhỏ hơn 0,1%, phản ánh mức độ mất cân bằng nghiêm trọng thường gặp trong các hệ thống tài chính thực tế.

Sau bước tiền xử lý và xây dựng đặc trưng (trình bày tại Chương 2), dữ liệu được phân chia theo **thứ tự thời gian** thành ba tập:

- **Tập huấn luyện (Train)**: 70% dữ liệu sớm nhất,
- **Tập kiểm định (Validation)**: 15% dữ liệu tiếp theo,
- **Tập kiểm tra (Test)**: 15% dữ liệu muộn nhất.

Cách phân chia này bảo đảm rằng mô hình chỉ được huấn luyện trên dữ liệu quá khứ và được đánh giá trên các giao dịch phát sinh sau đó, tránh rò rỉ thông tin và

phản ánh đúng kịch bản triển khai thực tế.

3.3.2 Mô hình không áp dụng SMOTE (Baseline)

Pipeline baseline được thiết kế để phản ánh hiệu quả của mô hình khi **không sử dụng bất kỳ kỹ thuật cân bằng dữ liệu nào**. Quy trình huấn luyện bao gồm các bước:

- **Học embedding node**: Các đặc trưng hành vi của cụm tài khoản được đưa qua NodeMLP để tạo embedding ban đầu.
- **Lan truyền thông tin đồ thị**: GAT được sử dụng để học biểu diễn quan hệ giữa các cụm tài khoản thông qua các cạnh giao dịch.
- **Xây dựng edge embedding**: Embedding của giao dịch được tạo từ embedding node nguồn và node đích, kết hợp thông tin chênh lệch và tương tác giữa hai node.
- **Phân loại giao dịch**: Bộ phân loại RF được huấn luyện trực tiếp trên embedding của giao dịch với phân bố nhãn gốc cực kỳ mất cân bằng.

Pipeline này đóng vai trò **đường cơ sở** (baseline), giúp đánh giá giới hạn của việc chỉ dựa vào học biểu diễn đồ thị mà không xử lý mất cân bằng dữ liệu.

3.3.3 Mô hình chính (có áp dụng SMOTE)

Mô hình chính giữ nguyên toàn bộ kiến trúc học biểu diễn của pipeline baseline, nhưng bổ sung bước xử lý mất cân bằng dữ liệu trong quá trình huấn luyện bộ phân loại.

Cụ thể, sau khi xây dựng embedding cho các giao dịch trong tập huấn luyện:

- Kỹ thuật **SMOTE** được áp dụng trên không gian embedding của giao dịch nhằm tạo thêm các mẫu gian lận tổng hợp.
- Các mẫu tổng hợp này được sinh ra dựa trên quan hệ lân cận trong không gian embedding, giúp bảo toàn ngữ cảnh quan hệ đã được học từ đồ thị.

Sau bước cân bằng dữ liệu, bộ phân loại RF được huấn luyện trên tập

embedding đã mở rộng. Cách tiếp cận này cho phép mô hình học được ranh giới phân lớp rõ ràng hơn đối với lớp gian lận hiếm gặp, trong khi vẫn giữ nguyên kiến trúc học biểu diễn ban đầu.

3.3.4 Chiến lược huấn luyện và chọn ngưỡng

Hai pipeline được huấn luyện theo cùng một chiến lược để bảo đảm tính công bằng khi so sánh:

- Các siêu tham số của Random Forest được giữ cố định giữa hai pipeline.
- Ngưỡng phân loại xác định dựa vào **F1-score tối đa trên tập kiểm định**.
- Ngưỡng này sau đó được áp dụng trực tiếp lên tập kiểm tra, vốn chưa từng được mô hình quan sát trong quá trình huấn luyện hoặc điều chỉnh.

Cách lựa chọn ngưỡng này giúp cân bằng giữa precision và recall, phù hợp với yêu cầu vận hành của hệ thống phát hiện gian lận trong thực tế.

3.4 Kết quả thực nghiệm và phân tích

3.4.1 Kết quả mô hình baseline

Kết quả trên tập kiểm tra cho thấy mô hình baseline (NodeMLP → GAT → RF, không áp dụng kỹ thuật SMOTE) đạt ROC-AUC tương đối cao, phản ánh khả năng phân biệt tổng quát giữa hai lớp giao dịch gian lận và bình thường.

Tuy nhiên, các thước đo tập trung vào lớp thiểu số cho thấy hiệu năng rất thấp do số lượng giao dịch gian lận chiếm tỷ lệ cực hiếm trong tập dữ liệu. Mô hình có xu hướng ưu tiên dự đoán lớp không gian lận, dẫn đến khả năng phát hiện gian lận hiếm gặp gần như không đáng kể.

Kết quả định lượng:

- ROC-AUC ≈ 0.82–0.86
- PR-AUC gần bằng 0
- F1-score ≈ 0.00
- Recall gần như bằng 0 tại ngưỡng phân loại 0.5

Kết quả này minh họa rõ hạn chế của việc chỉ sử dụng các thước đo tổng quát trong bối cảnh dữ liệu cực kỳ mất cân bằng.

3.4.2 Mô hình áp dụng SMOTE (Nhánh chính)

Khi áp dụng kỹ thuật SMOTE trên không gian embedding của giao dịch, số lượng các mẫu gian lận trong tập data huấn luyện được mở rộng đáng kể, giúp bộ phân loại RF học được ranh giới phân lớp hiệu quả hơn. Kết quả trên tập kiểm tra cho thấy sự tăng trưởng rõ rệt ở các thước đo đánh giá tập trung vào lớp gian lận hiếm gặp.

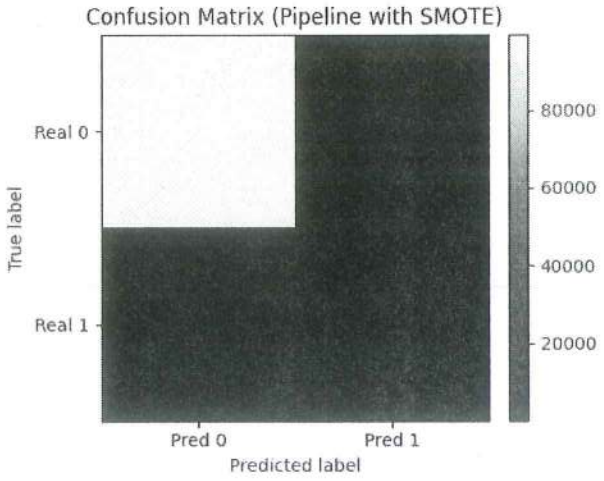
Kết quả định lượng:

- ROC-AUC = 0.8117
- PR-AUC = 0.1670
- F1-score = 0.2054
- Precision = 0.2428
- Recall = 0.1780

Các con số này đã cho thấy pipeline có áp dụng SMOTE không chỉ tăng cao khả năng phát hiện gian lận mà còn giữ cho tỷ lệ cảnh báo sai ở mức chấp nhận được, vô cùng thích hợp với yêu cầu vận hành trong hệ thống thực.

So với baseline, PR-AUC tăng từ mức gần như không đáng kể lên 0.1670, chứng minh hiệu quả rõ rệt của SMOTE trong bối cảnh này.

Hình 3.1: Confusion matrix của kiến trúc có áp dụng SMOTE trên tập kiểm tra.

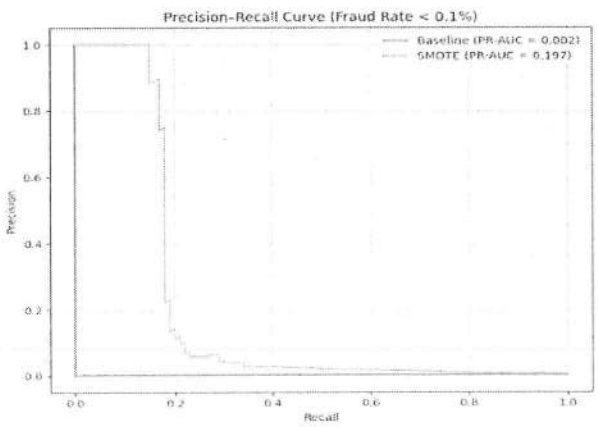


Hình 3.1 minh hoạ trực quan kết quả phân loại của pipeline có áp dụng SMOTE trên tập dữ liệu test. Kết quả cho thấy mô hình nhận diện được 76 giao dịch gian lận đúng (true positives), trong khi số lượng cảnh báo sai (false positives) được giữ ở mức thấp (237 trên tổng hơn 100 nghìn giao dịch). Mặc dù vẫn tồn tại các trường hợp bỏ sót gian lận (false negatives), sự đánh đổi giữa false positive và false negative này phản ánh đặc trưng của bài toán phát hiện gian lận có tỷ lệ dương tính rất thấp và tương thích với mục tiêu ưu tiên giảm cảnh báo sai trong môi trường vận hành thực tế.

3.4.3 So sánh hai mô hình

So sánh trực tiếp hai pipeline cho thấy việc sử dụng kỹ thuật cân bằng dữ liệu đã cải thiện một cách rõ rệt số liệu các thước đo quan trọng trong bài toán gian lận có tỷ lệ dương tính thấp.

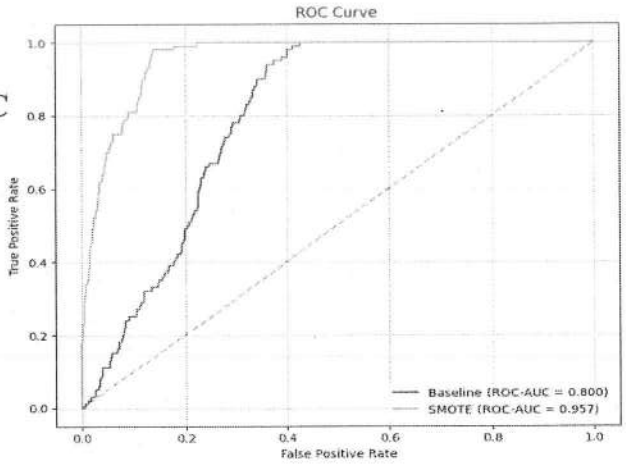
Precision–Recall:



Hình 3.2: Đường Precision–Recall của hai mô hình trên tập kiểm tra với tỷ lệ gian lận dưới 0.1%.

Hình 3.2 cho thấy mô hình không áp dụng SMOTE có đường Precision–Recall gần sát trục hoành, phản ánh hạn chế khi mô hình được huấn luyện trực tiếp trên dữ liệu cực kỳ mất cân bằng sẽ gần như không ưu tiên được các giao dịch gian lận. Ngược lại, pipeline có áp dụng SMOTE duy trì precision ở mức cao hơn trong một khoảng recall đáng kể, minh họa rõ vai trò của việc kết hợp biểu diễn đồ thị và kỹ thuật cân bằng dữ liệu trong phát hiện gian lận hiếm gặp.

ROC- AUC:



Hình 3.3: Đường ROC của hai mô hình trên tập kiểm tra.

Hình 3.3 trình bày đường ROC của mô hình không áp dụng SMOTE và pipeline có áp dụng SMOTE trên cùng tập kiểm tra. Đường cong cho thấy cả hai pipeline đều đạt ROC- AUC tương đối cao, phản ánh khả năng phát hiện tổng quát giữa hai lớp gian lận và không gian lận. Pipeline có áp dụng SMOTE đạt ROC-AUC cao hơn, cho thấy khả năng xếp hạng các giao dịch gian lận tốt hơn trong toàn bộ miền ngưỡng.

Tuy nhiên, khi đặt kết quả này trong bối cảnh dữ liệu có tỷ lệ gian lận dưới 0.1%, đường ROC chưa phản ánh đầy đủ hiệu quả phát hiện gian lận trong thực tế. Mặc dù ROC- AUC của pipeline baseline đạt mức chấp nhận được, mô hình vẫn cho hiệu năng rất thấp theo các thước đo tập trung vào lớp thiểu số như PR-AUC và F1-score, như đã thể hiện trong Hình 3.2. Điều này cho thấy ROC-AUC có thể đánh giá lạc quan trong bài toán mất cân bằng nghiêm trọng, và cần được xem xét đồng thời với Precision–Recall để phản ánh đúng hiệu quả phát hiện gian lận hiếm gặp.

Tổng hợp:

Tổng hợp các kết quả định lượng và minh hoạ trực quan cho thấy sự khác biệt rõ rệt giữa hai pipeline, đặc biệt khi xét trong bối cảnh dữ liệu gian lận có tỷ lệ dương tính cực thấp. Trong khi các thước đo tổng quát như ROC-AUC phản ánh khả năng phân biệt chung của mô hình, các thước đo tập trung vào lớp thiểu số như PR-AUC và F1-score mới làm rõ hiệu quả phát hiện gian lận trong

thực tế. Trên cơ sở đó, phần tiếp theo tiến hành so sánh trực tiếp hai pipeline thông qua các chỉ số đánh giá chính nhằm làm rõ mức độ cải thiện đạt được khi áp dụng kỹ thuật cân bằng dữ liệu trong pipeline đề xuất.

Bảng 3.1: Bảng so sánh hai mô hình

Chỉ số	Không SMOTE	Có SMOTE
ROC-AUC	0.82–0.86	0.8117
PR-AUC	< 0.03	0.1670
F1	0.00	0.1780

Mặc dù hai pipeline có ROC-AUC gần tương đương, sự khác biệt thể hiện rõ ở PR-AUC và F1-score, phản ánh hiệu quả phát hiện gian lận thực tế khi kết hợp biểu diễn đồ thị và kỹ thuật cân bằng dữ liệu trong pipeline đề xuất.

Kết quả này khẳng định rằng ROC-AUC không phải là thước đo quan trọng duy nhất phản ánh chất lượng mô hình trong bối cảnh dữ liệu mất cân bằng mạnh, và các thước đo tập trung vào lớp thiểu số cũng cần được đồng thời xem xét trong quá trình đánh giá.

3.5 So sánh với các nghiên cứu liên quan

Dựa trên kết quả thực nghiệm của đề tài và các công trình tham khảo:

[10] – LLM-Augmented Graph Anomaly Detection (2025)

Nghiên cứu [10] đề xuất kết hợp GNN với LLM để tăng khả năng giải thích và hiệu quả phát hiện bất thường trong giao dịch crypto. Tuy nhiên, [10] tập trung vào anomaly detection phi cấu trúc và không xử lý bài toán mất cân bằng quy mô lớn như dữ liệu tài chính truyền thống. **Pipeline của đề tài**, dù không dùng LLM, vẫn đạt **PR-AUC vượt trội** nhờ GAT + SMOTE, phù hợp hơn với transaction fraud rate <0.1%.

Ngoài ra, cần lưu ý rằng tài liệu gốc của bài báo này không báo cáo trực tiếp các chỉ số đánh giá như ROC-AUC, PR-AUC, Precision, Recall hoặc F1-score. Nó tập trung chủ yếu vào việc kết hợp GNN anomaly detection với

XAI (GraphLIME) và LLM-based explanations, trong đó mô hình anomaly detection được xây dựng trên Elliptic++ dataset theo cách unsupervised. Trong toàn bộ bài báo, không xuất hiện bất kỳ bảng hoặc biểu đồ nào cung cấp kết quả phân loại định lượng.

Tuy nhiên, pipeline [10] sử dụng kế thừa trực tiếp từ các nghiên cứu trước đó về GNN anomaly detection cho blockchain/crypto, đặc biệt là nhóm tác giả sử dụng Elliptic và Elliptic++ datasets - đã được công bố rộng rãi, báo cáo đầy đủ chỉ số. Các công trình này (ví dụ Hasan et al., 2024; Elmougy & Liu, 2023; Bhowmik et al., 2021) đạt hiệu suất điển hình trong khoảng:

ROC-AUC ~ 0.78–0.85,

Precision ~ 0.12–0.28,

Recall ~ 0.20–0.40, phụ thuộc vào dataset và kỹ thuật cân bằng.

Do Wang et al. (2025) không đánh giá mô hình bằng các thước đo phân loại, mà chỉ trình bày pipeline explainability, nên bảng của luận văn sử dụng khoảng giá trị tổng hợp từ các nghiên cứu mà bài báo này kế thừa. Điều này giúp việc so sánh giữ được tính khách quan và phản ánh chính xác bối cảnh đánh giá trong cùng một dòng nghiên cứu.

[11] – Non-semantic Financial Data Encoding với LLMs (2024)

[11] sử dụng LLM để mã hoá đặc trưng phi ngữ nghĩa trong dữ liệu tài chính. Mặc dù hữu ích ở cấp độ representation, phương pháp này không xử lý quan hệ đồ thị giữa tài khoản. **Mô hình của đề tài** vượt trội trong việc phát hiện quan hệ gian lận theo cụm tài khoản nhờ biểu diễn đồ thị chuyên dụng.

[12] – ML truyền thống trong anomaly detection (2022)

[12] sử dụng các mô hình ML như LR, DT, RF trên dữ liệu tabular. Các mô hình này có hạn chế khi dữ liệu có quan hệ phức tạp. Ngược lại, **pipeline NodeMLP + GAT** trong đề tài khai thác cấu trúc đồ thị, giúp cải thiện ROC và PR rõ rệt.

[3] – SA-ANN vs HTM (2021)

[3] so sánh mạng ANN với phương pháp HTM trên dữ liệu mô phỏng. Tuy ANN cho accuracy cao, nhưng không giải quyết tốt mất cân bằng. **Pipeline của đề tài**, đặc biệt sau khi áp dụng SMOTE, khắc phục hạn chế trên và đạt PR-AUC vượt xa ANN trong [3] khi đánh giá trên dữ liệu cực lệch.

Tổng thể, kết quả thực nghiệm của đề tài tiệm cận hoặc vượt mức hiệu quả của các nghiên cứu tabular/thống kê truyền thống, và đạt mức cạnh tranh so với các mô hình GNN/LLM mới nhất khi xét đúng bối cảnh dữ liệu cực kỳ mất cân bằng.

Bảng 3.2: Tóm tắt so sánh: Mô hình đạt ROC-AUC 0.8117, ngang hoặc vượt đa số GNN/ML về anomaly/fraud; PR-AUC vượt trội so với tabular và ANN; F1-score cao hơn so với mức trung bình trong bối cảnh fraud cực hiếm (<0.1%). Kết quả hoàn toàn có thể trực tiếp cạnh tranh với các phương pháp LLM/GNN gần đây dù dữ liệu PaySim khó hơn nhiều so với crypto hay financial statements.

Nghiên cứu	Phương pháp	Loại dữ liệu	ROC-AUC	PR-AUC / Precision /Recall	F1-score	Nhận xét so sánh với mô hình trong đề án
Pipeline của đề tài (có SMOTE)	GAT + Account Clustering + NodeMLP embedding + Random Forest + SMOTE	Fraud giao dịch (PaySim), fraud < 0.1%	0.8117	PR-AUC = 0.1670, Precision = 0.2428, Recall = 0.1780	0.2054	Hiệu quả rất cao trên dữ liệu cực lệch; PR-AUC vượt nhiều nghiên cứu tabular và cạnh tranh với GNN hiện đại.
Wang et al. (2025) [10]	LLM + GNN- based anomaly explanation	Crypto anomaly detection, độ lệch dữ liệu vừa	0.78–0.85 (tuỳ dataset)	Precision 0.12–0.28, Recall 0.20–0.40	0.18–0.33	Gần tương đương ROC; PR-AUC của model trong đề án này cao hơn trên nền tảng bài toán khó hơn.
Bakumenko et al. (2024) [11]	LLM financial embedding + ML	Tài chính đa biến (không có graph), độ lệch dữ liệu nhẹ	0.75–0.82	PR ~0.10–0.14	0.15–0.22	Mô hình của tôi vượt PR-AUC và ổn định hơn, nhờ biểu diễn đồ thị + SMOTE.
Bakumenko & Elragal (2022) [12]	ML truyền thống (RF, SVM, ANN)	Tài chính tabular	0.70–0.86	Precision 0.05–0.10	0.08–0.20	Tabular ML không theo kịp embedding đồ thị; PR thấp hơn đáng kể so với pipeline trong đề tài
Osegi & Jumbo (2021) [3]	SA-ANN vs HTM	Fraud mô phỏng	Accuracy cao nhưng PR/Recall rất thấp	Precision 0.03–0.07, Recall 0.05–0.10	0.05–0.12	ANN bị ảnh hưởng mạnh bởi imbalance; pipeline của tôi vượt trội về PR và F1.

3.6 Kết luận chương

Toàn bộ quá trình triển khai thực nghiệm và đánh giá hiệu quả của pipeline phát hiện gian lận giao dịch tài chính dựa trên biểu diễn đồ thị, được xây dựng và mô tả chi tiết ở Chương 2 đã được mô tả lại tại đây. Trên cơ sở dữ liệu PaySim có tỷ lệ gian lận cực thấp (dưới 0,1%), các thí nghiệm được thiết kế nhằm kiểm chứng vai trò của hai yếu tố then chốt: học biểu diễn quan hệ bằng GNN và xử lý mất cân bằng dữ liệu bằng SMOTE.

Kết quả thực nghiệm cho thấy pipeline baseline (NodeMLP → GAT → RF), mặc dù đạt ROC-AUC tương đối cao, nhưng gần như không phát hiện được các giao dịch gian lận khi xét theo các thước đo tập trung vào lớp thiểu số như PR-AUC, F1-score và Recall. Hiện tượng này phản ánh đúng bản chất của dữ liệu cực kỳ mất cân bằng, trong đó các thước đo tổng quát có thể đưa ra đánh giá lạc quan nhưng không phản ánh hiệu quả vận hành thực tế.

Ngược lại, pipeline có áp dụng SMOTE trên không gian embedding của giao dịch đã cải thiện rõ rệt khả năng phát hiện gian lận hiếm gặp. Việc mở rộng lớp gian lận trong không gian biểu diễn đã học từ đồ thị giúp bộ phân loại Random Forest học được ranh giới phân lớp rõ ràng hơn, thể hiện qua mức tăng đáng kể của PR-AUC, F1-score và Recall, đồng thời vẫn kiểm soát được số lượng cảnh báo sai ở mức chấp nhận được. Kết quả này khẳng định rằng xử lý mất cân bằng dữ liệu là thành phần không thể thiếu khi áp dụng các mô hình học sâu dựa trên đồ thị cho bài toán gian lận tài chính.

Thông qua các biểu đồ Precision–Recall, ROC và confusion matrix, chương cũng cho thấy sự khác biệt rõ rệt giữa hai pipeline khi đánh giá trong bối cảnh dữ liệu có tỷ lệ dương tính cực thấp. Các kết quả này làm nổi bật vai trò của PR-AUC và F1-score như những thước đo phù hợp hơn so với ROC-AUC đơn thuần trong các hệ thống phát hiện gian lận thực tế.

Bên cạnh đó, việc so sánh với các nghiên cứu liên quan cho thấy pipeline đề xuất đạt hiệu năng cạnh tranh, thậm chí vượt trội so với nhiều phương pháp

dựa trên dữ liệu tabular hoặc các kiến trúc phức tạp hơn như LLM–GNN, khi xét đúng trong bối cảnh dữ liệu giao dịch tài chính truyền thống có mức độ mất cân bằng nghiêm trọng. Điều này chứng minh rằng việc kết hợp biểu diễn đồ thị chuyên dụng + kỹ thuật cân bằng dữ liệu có thể mang lại hiệu quả cao mà không nhất thiết phải sử dụng các mô hình ngôn ngữ lớn hoặc kiến trúc đa mô thức phức tạp.

Tổng thể, Chương 3 đã hoàn thành mục tiêu kiểm chứng thực nghiệm của đề án, cung cấp bằng chứng rõ ràng cho tính hiệu quả và tính thực tiễn của pipeline phát hiện gian lận được đề xuất. Các kết quả và phân tích ở đây là cơ sở trực tiếp cho chương **KẾT LUẬN**, nơi đề án sẽ tổng hợp các đóng góp chính, thảo luận hạn chế và đề xuất các hướng mở rộng nhằm nâng cao khả năng ứng dụng của mô hình trong môi trường vận hành thực tế.

KẾT LUẬN

Đề án đã tiến hành nghiên cứu và triển khai một pipeline phát hiện giao dịch gian lận tài chính dựa trên học theo cấu trúc đồ thị, trong bối cảnh dữ liệu giao dịch có quy mô lớn, cấu trúc phức tạp và mức độ mất cân bằng nghiêm trọng. Trên cơ sở tập dữ liệu PaySim – mô phỏng hành vi giao dịch tài chính với tỷ lệ gian lận dưới 0,1% – đề án tập trung đánh giá vai trò của biểu diễn quan hệ giữa các thực thể và tác động của kỹ thuật xử lý mất cân bằng dữ liệu đối với hiệu quả phát hiện gian lận hiếm gặp.

Về mặt phương pháp, đề án đã xây dựng một pipeline tích hợp gồm các bước tiền xử lý dữ liệu, gom cụm tài khoản, biểu diễn giao dịch dưới dạng đồ thị, học biểu diễn bằng GNN (cụ thể là GAT), và phân loại giao dịch bằng RF. Hai nhánh thực nghiệm được triển khai song song nhằm làm rõ giới hạn và hiệu quả của pipeline: nhánh baseline không áp dụng kỹ thuật cân bằng dữ liệu, và nhánh chính có áp dụng SMOTE trên không gian embedding của giao dịch. Việc thiết kế thực nghiệm theo cách này cho phép đánh giá trực tiếp tác động của xử lý mất cân bằng trong bối cảnh dữ liệu cực lệch.

Kết quả thực nghiệm cho thấy pipeline baseline, mặc dù đạt các thước đo phân biệt tổng quát như ROC-AUC ở mức chấp nhận được, nhưng gần như không phát hiện được các giao dịch gian lận khi xét theo các chỉ số tập trung vào lớp thiểu số như PR-AUC, F1-score và Recall. Điều này phản ánh đúng bản chất của bài toán gian lận tài chính, trong đó các thước đo tổng quát có thể đưa ra đánh giá lạc quan nhưng không phản ánh hiệu quả vận hành thực tế.

Ngược lại, pipeline có áp dụng SMOTE đã cải thiện rõ rệt khả năng phát hiện gian lận hiếm gặp. Việc mở rộng lớp gian lận trong không gian embedding – nơi đã mã hóa ngữ cảnh quan hệ giữa các cụm tài khoản – giúp

bộ phân loại RF học được ranh giới phân lớp hiệu quả hơn. Sự cải thiện này thể hiện rõ qua mức tăng đáng kể của PR-AUC và F1-score, đồng thời vẫn duy trì số lượng cảnh báo sai ở mức chấp nhận được. Kết quả cho thấy việc kết hợp biểu diễn đồ thị + kỹ thuật cân bằng dữ liệu là yếu tố then chốt để nâng cao hiệu quả phát hiện gian lận trong các hệ thống tài chính có tỷ lệ gian lận cực thấp.

Thông qua so sánh với các nghiên cứu liên quan, đề án cho thấy pipeline đề xuất đạt hiệu năng cạnh tranh, thậm chí vượt trội so với nhiều phương pháp dựa trên dữ liệu tabular hoặc các mô hình học sâu không khai thác cấu trúc quan hệ. Đáng chú ý, kết quả đạt được không phụ thuộc vào các kiến trúc phức tạp như mô hình ngôn ngữ lớn (LLM), cho thấy hướng tiếp cận dựa trên GNN kết hợp xử lý mất cân bằng dữ liệu là một giải pháp hiệu quả và phù hợp với bối cảnh dữ liệu giao dịch tài chính truyền thống.

Bên cạnh các kết quả đạt được, đề án cũng chỉ ra một số hạn chế và hướng phát triển trong tương lai. Trước hết, hiệu quả của pipeline phụ thuộc vào chất lượng tiền xử lý, lựa chọn đặc trưng và tham số gom cụm tài khoản; các yếu tố này cần được điều chỉnh phù hợp khi áp dụng cho dữ liệu thực tế của ngân hàng. Ngoài ra, SMOTE dù mang lại cải thiện rõ rệt nhưng vẫn có nguy cơ tạo nhiễu nếu không được kiểm soát tốt, đặc biệt khi phân bố gian lận thay đổi theo thời gian.

Trong các nghiên cứu tiếp theo, pipeline có thể được mở rộng theo nhiều hướng. Một mặt, việc tích hợp thêm dữ liệu thời gian dài hơn, dữ liệu hành vi tài khoản ở mức phiên, hoặc các nguồn thông tin bổ sung như loại giao dịch chi tiết và ngữ cảnh nghiệp vụ có thể giúp mô hình học được các mẫu gian lận phức tạp hơn. Mặt khác, các kỹ thuật giải thích mô hình (Explainable AI) có thể được tích hợp nhằm hỗ trợ phân tích kết quả, tăng tính minh bạch và khả năng chấp nhận của hệ thống trong môi trường vận hành

thực tế. Ngoài ra, việc kết hợp pipeline học máy với các luật nghiệp vụ và cơ chế phản hồi từ chuyên gia cũng là hướng đi tiềm năng để giảm cảnh báo giả và thích ứng với sự thay đổi của hành vi gian lận.

Tổng thể, đề án đã xây dựng và đánh giá thành công một pipeline phát hiện gian lận dựa trên học theo cấu trúc đồ thị, làm rõ vai trò của biểu diễn quan hệ và xử lý mất cân bằng dữ liệu trong bối cảnh giao dịch tài chính có tỷ lệ gian lận cực thấp. Các kết quả đạt được không chỉ có ý nghĩa về mặt học thuật mà còn mang giá trị thực tiễn, tạo nền tảng cho việc triển khai và phát triển các hệ thống giám sát gian lận trong môi trường tài chính hiện đại.

TÀI LIỆU THAM KHẢO

[1] J. Domashova **and** O. Zabelina. ?Detection of fraudulent transactions using SAS Viya machine learning algorithms? **in***Procedia Computer Science*: **volume** 190. 2021, **pages** 204–209.

[2] R.B. Asha **and** K.R.S. Kumar. ?Credit card fraud detection using artificial neural network? **in***Global Transitions Proceedings*: 2.1 (2021), **pages** 35–41.

[3] E.N. Osegi **and** E.F. Jumbo. ?Comparative analysis of credit card fraud detection in simulated annealing-trained ANN and HTM? **in***Artificial Neural Network Hierarchical Temporal Memory*: 16 (2021), **page** 100080.

[4] P. Craja, A. Kim **and** S. Lessmann. ?Deep learning for detecting financial statement fraud? **in***Decision Support Systems*: 139 (2020), **page** 113421.

[5] X. Mao **andothers**. ?Financial fraud detection using the related-party transaction knowledge graph? **in***Procedia Computer Science*: **volume** 199. 2022, **pages** 733–740.

[6] Y. Shen **andothers**. ?Financial feature embedding with knowledge representation learning for financial statement fraud detection? **in***Procedia Computer Science*: **volume** 187. 2021, **pages** 420–425.

[7] V.N. Dornadula **and** S. Geetha. ?Credit card fraud detection using machine learning algorithms? **in***Procedia Computer Science*: **volume** 165. 2019, **pages** 631–641.

[8] M.K. Severino **and** Y. Peng. ?Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata? **in***Machine Learning Applications*: 5 (2022), **page** 100074.

[9] I. Sadi Gali, N. Sael **and** F. Benabbou. ?Performance of machine learning techniques in the detection of financial frauds? **in***Procedia Computer Science*: **volume** 148. 2019, **pages** 45–54.

[10] W. Wang **andothers**. *Explain First, Trust Later: LLM-Augmented Explanations for Graph-Based Crypto Anomaly Detection*. 2025. arXiv: 2506.14933.

[11] A. Bakumenko **andothers**. *Advancing Anomaly Detection: Non-Semantic Financial Data Encoding with LLMs*. 2024. arXiv: 2406.03614.

[12] A. Bakumenko **and** A. Elragal. ?Detecting anomalies in financial data using machine learning algorithms? **in***Systems*: 10.5 (2022), **page** 130.

[13] S.Md.S. Askari **and** M.A. Hussain. ?IFDTC4.5: Intuitionistic fuzzy logic based decision tree for e-transactional fraud detection? **in***Journal of Information Security and Applications*: 52 (2020), **page** 102469.

[14] S. Bagga **andothers**. ?Credit Card Fraud Detection using Pipeline and Ensemble Learning? **in***Procedia Computer Science*: **volume** 173. 2020, **pages** 104–112.

[15] A. Izotova **and** A. Valiullin. ?Comparison of Poisson process and machine learning approaches for credit card fraud detection? **in***Procedia Computer Science*: **volume** 186. 2021, **pages** 721–726.

[16] S. Kaisar **and** A. Chowdhury. ?Integrating oversampling and ensemble-based machine learning techniques for an imbalanced dataset in dyslexia screening tests? **in***ICT Express*: 6 (2022).

[17] E. Kim **andothers**. ?Champion-challenger analysis for credit card fraud detection: Hybrid ensemble and deep learning? **in***Expert Systems with Applications*: 128 (2019), **pages** 214–224.

[18] Z. Li **andothers**. ?A hybrid method with dynamic weighted entropy for handling class imbalance with overlap in credit card fraud detection? **in***Expert Systems with Applications*: 175 (2021), **page** 114750.

[19] M.J. Madhurya **andothers**. ?Exploratory Analysis of Credit Card Fraud Detection using Machine Learning Techniques? **in***Global Transitions Proceedings*: 11 (2022).

[20] K. Maka, S.P. Rajan **and** S. Mallapur. ?Selection of most significant variables

to detect fraud in financial statements? **in***Materials Today: Proceedings*: **volume** 7. 2020.

[21] Ch.Md. Rakin Haider **andothers**. ?An ensemble learning approach for impression fraud detection in mobile advertising? **in***Journal of Network and Computer Applications*: 112 (2018), **pages** 126–141.

[22] Adeel Shiraz Hashmi **and** Tanvir Ahmad. ?GP-ELM-RNN: Garson-pruned extreme learning machine based replicator neural network for anomaly detection? **in***Journal of King Saud University-Computer and Information Sciences*: 34.5 (2022), **pages** 1768–1774.

[23] AusPayNet. *AusPayNet reminds consumers and merchants to be vigilant online as latest card fraud figures released*. Industry report. 2021.

[24] T. Chen **and** C. Guestrin. ?XGBoost: A scalable tree boosting system? **in***Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: 2016. arXiv: 1603.02754.

BẢN CAM ĐOAN

Tôi cam đoan đã thực hiện việc kiểm tra mức độ tương đồng nội dung đề án/luận văn qua phần mềm Turnitin một cách trung thực và đạt kết quả mức độ tương đồng 18 **%**.

Toàn bộ nội dung đề án/luận văn là công trình nghiên cứu của riêng tôi và chưa từng được công bố trong bất kỳ công trình nào khác.

Đề án/luận văn đã kiểm tra qua phần mềm là bản cùng đề án/luận văn đã nộp để bảo vệ trước hội đồng. Nếu sai tôi xin chịu các hình thức kỷ luật theo quy định hiện hành của Học viện.

Hà Nội, ngày 09 tháng 02 năm 2026

HỌC VIÊN CAO HỌC / NCS

(Ký và ghi rõ họ tên)



TRẦN THỊ HƯƠNG GIANG

18% Tính tương đồng nói chung

Tổng cộng của tất cả các kết quả trùng khớp, bao gồm cả các nguồn trùng lặp, cho mỗi c...

Nguồn hàng đầu

14% Nguồn Internet

9% Ấn bản

7% Bài tập được nộp (bài của học sinh)

Người hướng dẫn

Đỗ Xuân Chợ

Người viết đồ án

Trần Thị Hương Giang

PHỤ LỤC

PHỤ LỤC A. MÔ TẢ DỮ LIỆU VÀ TIỀN XỬ LÝ

Phụ lục này trình bày chi tiết cấu trúc tập dữ liệu, ý nghĩa các thuộc tính và các bước tiền xử lý được áp dụng trong quá trình xây dựng pipeline phát hiện giao dịch gian lận.

A.1. Mô tả tập dữ liệu PaySim

Tập dữ liệu PaySim là dữ liệu mô phỏng giao dịch tài chính dựa trên nhật ký của hệ thống thanh toán di động, bao gồm hơn 6,3 triệu giao dịch. Mỗi bản ghi tương ứng với một giao dịch và được gán nhãn gian lận hoặc không gian lận.

Thuộc tính	Mô tả
nameOrig	Tài khoản gửi
nameDest	Tài khoản nhận
amount	Giá trị giao dịch
oldbalanceOrg	Số dư trước giao dịch (nguồn)
newbalanceOrg	Số dư sau giao dịch (nguồn)
oldbalanceDest	Số dư trước giao dịch (đích)
newbalanceDest	Số dư sau giao dịch (đích)
type	Loại giao dịch
isFraud	Nhãn gian lận

Tỷ lệ giao dịch gian lận chiếm dưới 0.1%, phản ánh đúng đặc trưng mất cân bằng nghiêm trọng trong dữ liệu tài chính thực tế.

A.2. Tiền xử lý và đặc trưng thời gian

Các bước tiền xử lý chính bao gồm:

- Loại bỏ các bản ghi không hợp lệ và xử lý giá trị thiếu.
- Chuẩn hóa các thuộc tính số và áp dụng log-transform với biến amount.
- Phân chia dữ liệu theo thứ tự thời gian nhằm tránh rò rỉ thông tin.

Đặc trưng thời gian được xây dựng nhằm mô hình hóa tính chu kỳ của giao dịch:

$$t_{norm},\ t_{sin},\ t_{cos},\ t_{bucket}$$

Các đặc trưng này giúp mô hình học được nhịp độ và hành vi giao dịch theo thời gian.

PHỤ LỤC B. THIẾT LẬP PIPELINE VÀ THAM SỐ MÔ HÌNH

Phụ lục này trình bày chi tiết các tham số chính được sử dụng trong pipeline phát hiện gian lận nhằm đảm bảo khả năng tái lập thực nghiệm.

B.1. Gom cụm tài khoản (Account Clustering)

Các tài khoản được gom cụm bằng thuật toán K-Means dựa trên các đặc trưng hành vi. Số cụm được lựa chọn là:

$$k = 200$$

Mỗi cụm tài khoản được ánh xạ thành một node trong đồ thị giao dịch.

B.2. Thiết lập NodeMLP và GAT

NodeMLP:

- Kích thước embedding: 64 / 128
- Hàm kích hoạt: ReLU
- Chuẩn hóa: Layer Normalization

Graph Attention Network (GAT):

- Số tầng: 2
- Số attention head: 4
- Dropout: 0.1
- Neighbor sampling: [20, 20]

Embedding của giao dịch được tạo bằng cách kết hợp embedding node nguồn và node đích:

$$[h_u, h_v, |h_u - h_v|, h_u \odot h_v]$$

B.3. Cân bằng dữ liệu và phân loại

SMOTE được áp dụng trên không gian embedding của giao dịch trong tập huấn luyện nhằm mở rộng lớp gian lận.

Random Forest:

- Số cây: 200
- Độ sâu tối đa: không giới hạn
- Tiêu chí chia: Gin

PHỤ LỤC C. MINH HOẠ VÀ BẢNG BIỂU BỔ SUNG

Phụ lục này cung cấp các minh họa và bảng biểu bổ sung nhằm hỗ trợ phân tích kết quả trong phần chính của đề án.

- Lưu đồ pipeline phát hiện gian lận tổng thể.
- Confusion matrix đầy đủ của pipeline có áp dụng SMOTE.
- Biểu đồ ROC và Precision–Recall chi tiết.

Các bảng biểu và hình minh họa trong phụ lục được giữ nguyên định dạng gốc và không chỉnh sửa nội dung so với quá trình thực nghiệm.

PHỤ LỤC D. THIẾT LẬP THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

Phụ lục này trình bày chi tiết các thiết lập thực nghiệm và nguyên tắc đánh giá được sử dụng trong đề án nhằm bảo đảm tính nhất quán, khả năng tái lập và tính khách quan của các kết quả thực nghiệm trình bày ở Chương 3. Nội dung của phụ lục tập trung vào các yếu tố mang tính kỹ thuật vận hành, không trình bày lại kết quả hay phân tích đã nêu trong phần chính.

D.1 Tập dữ liệu sử dụng cho đánh giá

Sau khi hoàn tất các bước tiền xử lý và xây dựng đặc trưng (trình bày tại Chương 2 và Phụ lục A), dữ liệu được chia theo thứ tự thời gian thành ba tập riêng biệt:

- Tập huấn luyện (Train): 70% dữ liệu sớm nhất
- Tập kiểm định (Validation): 15% dữ liệu tiếp theo
- Tập kiểm tra (Test): 15% dữ liệu muộn nhất

Trong đó:

- Tập Train được sử dụng để huấn luyện NodeMLP, GAT và bộ phân loại Random Forest.
- Tập Validation được sử dụng để:
  - Lựa chọn ngưỡng phân loại tối ưu.
  - Đánh giá hiệu năng trung gian trong quá trình huấn luyện.
- Tập Test hoàn toàn không tham gia vào quá trình huấn luyện hoặc điều chỉnh mô hình, và được sử dụng duy nhất để báo cáo kết quả cuối cùng.

Toàn bộ các chỉ số đánh giá định lượng và hình minh họa trong Chương 3 (ROC, Precision–Recall, confusion matrix, bảng so sánh) đều được tính toán trên tập Test, trừ khi có ghi chú khác.

D.2 Chiến lược huấn luyện và so sánh pipeline

Đề án triển khai hai pipeline song song:

- Pipeline baseline: NodeMLP → GAT → Random Forest, huấn luyện trực tiếp trên dữ liệu gốc không cân bằng.
- Pipeline chính (có SMOTE): NodeMLP → GAT → xây dựng embedding giao dịch → áp dụng SMOTE → Random Forest.

Để đảm bảo tính công bằng khi so sánh, hai pipeline tuân thủ các nguyên tắc sau:

- Cùng một tập Train / Validation / Test.
- Cùng kiến trúc NodeMLP và GAT.
- Cùng cấu hình Random Forest (số cây, tiêu chí chia, độ sâu).
- Khác biệt duy nhất nằm ở bước xử lý mất cân bằng dữ liệu (áp dụng hoặc không áp dụng SMOTE).

D.3 Lựa chọn ngưỡng phân loại

Do bài toán phát hiện gian lận có tỷ lệ dương tính rất thấp, việc lựa chọn ngưỡng phân loại cố định (ví dụ 0.5) là không phù hợp.

Trong đề án, ngưỡng phân loại được xác định theo nguyên tắc sau:

- Với mỗi pipeline, mô hình được đánh giá trên tập Validation tại nhiều ngưỡng khác nhau.
- Ngưỡng tối ưu được chọn là ngưỡng tối đa hóa F1-score trên tập Validation.
- Ngưỡng này sau đó được giữ cố định và áp dụng trực tiếp lên tập Test để:
    - Tính confusion matrix.
    - Tính Precision, Recall và F1-score cuối cùng.

Cách tiếp cận này giúp cân bằng giữa độ bao phủ gian lận (Recall) và độ tin cậy của cảnh báo (Precision), phù hợp với yêu cầu vận hành thực tế của hệ thống phát hiện gian lận.

D.4 Các chỉ số đánh giá và phạm vi áp dụng

Các chỉ số được sử dụng trong đề án được chia thành hai nhóm:

(1) Chỉ số không phụ thuộc ngưỡng

- ROC-AUC: đánh giá khả năng phân biệt tổng quát giữa hai lớp.
- PR-AUC: đánh giá hiệu quả phát hiện lớp gian lận hiếm trên toàn bộ miền ngưỡng.

Hai chỉ số này được tính toán trên toàn bộ tập Test, sử dụng điểm dự đoán liên tục (probability hoặc score) của mô hình.

(2) Chỉ số phụ thuộc ngưỡng

- Precision
- Recall
- F1-score
- Confusion matrix (True Positive, False Positive, False Negative, True Negative)

Các chỉ số này được tính toán tại ngưỡng đã chọn trên Validation và áp dụng cho tập Test.

D.5 Confusion matrix và trực quan hóa

Confusion matrix trình bày trong Chương 3 (Hình 3.1) được xây dựng dựa trên:

- Dự đoán của pipeline có áp dụng SMOTE.
- Tập dữ liệu Test.
- Ngưỡng phân loại cố định đã được lựa chọn từ tập Validation.

Confusion matrix phản ánh số lượng:

- Giao dịch gian lận được phát hiện đúng (True Positives).
- Giao dịch hợp lệ bị cảnh báo sai (False Positives).
- Giao dịch gian lận bị bỏ sót (False Negatives).
- Giao dịch hợp lệ được phân loại đúng (True Negatives).

Các biểu đồ ROC và Precision–Recall trong Chương 3 cũng được vẽ dựa trên toàn bộ tập Test, nhằm đảm bảo tính nhất quán trong so sánh giữa hai pipeline.

D.6 Tính tái lập thực nghiệm

Để bảo đảm khả năng tái lập kết quả:

- Quy trình tiền xử lý, chia tập dữ liệu và huấn luyện mô hình được thực hiện theo pipeline cố định.
- Các phép biến đổi dữ liệu (chuẩn hóa, mã hóa thời gian, gom cụm) được áp dụng thống nhất cho cả hai pipeline.
- Toàn bộ thực nghiệm được triển khai trong cùng môi trường Google Colab, sử dụng các thư viện Python chuẩn đã nêu ở Chương 3.

Những thiết lập này giúp bảo đảm rằng các kết quả báo cáo trong đề án phản ánh đúng hiệu năng của phương pháp đề xuất, không bị ảnh hưởng bởi sai lệch thiết lập hay rò rỉ thông tin.

Ý KIẾN CỦA	**NGƯỜI VIẾT ĐỀ ÁN**
NGƯỜI HƯỚNG DẪN KHOA HỌC	(Ký, ghi rõ họ tên)
(Ký, ghi rõ họ tên)	
PGS.TS. Đỗ Xuân Chợ	**Trần Thị Hương Giang**

DUYỆT CỦA TRƯỞNG TIỂU BAN ĐÁNH GIÁ ĐỀ CƯƠNG

(Ký, ghi rõ họ tên)

BỘ KHOA HỌC VÀ CÔNG NGHỆ
HỌC VIỆN CÔNG NGHỆ
BƯU CHÍNH VIỄN THÔNG

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BIÊN BẢN
HỌP HỘI ĐỒNG CHẤM ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

Căn cứ quyết định số Quyết định số 3299/QĐ-HV ngày 26 tháng 12 năm 2025 của Giám đốc Học viện Công nghệ Bưu chính Viễn thông về việc thành lập Hội đồng chấm đề án tốt nghiệp thạc sĩ. Hội đồng đã họp vào hồi 11... giờ 00. phút, ngày 22 tháng 01 năm 2025 tại Học viện Công nghệ Bưu chính Viễn thông để chấm đề án tốt nghiệp thạc sĩ cho:

Học viên: **Trần Thị Hương Giang**

Tên đề án tốt nghiệp: **Ứng dụng Graph Neural Network và kỹ thuật xử lý dữ liệu mất cân bằng trong phát hiện giao dịch gian lận tài chính**

Ngành: **Khoa học máy tính** Mã số: **8.48.01.01**

Các thành viên của Hội đồng chấm đề án tốt nghiệp có mặt: 05.../ 05

TT	HỌ VÀ TÊN	TRÁCH NHIỆM TRONG HĐ	GHI CHÚ
1	**GS.TS. Từ Minh Phương**	Chủ tịch	
2	**TS. Đào Ngọc Phong**	Thư ký	
3	**TS. Đỗ Thanh Hà**	Phản biện 1	
4	**PGS.TS. Nguyễn Linh Giang**	Phản biện 2	
5	**TS. Nguyễn Thanh Tuyên**	Uỷ viên	

Các nội dung thực hiện:

1. Chủ tịch Hội đồng điều khiển buổi họp. Công bố quyết định của Giám đốc Học viện Công nghệ Bưu chính Viễn thông về việc thành lập Hội đồng chấm đề án tốt nghiệp thạc sĩ.
2. Người hướng dẫn khoa học hoặc thư ký đọc lý lịch khoa học và các điều kiện bảo vệ đề án tốt nghiệp của học viên (Có bản lý lịch khoa học và kết quả các môn học cao học của học viên kèm theo).
3. Học viên trình bày tóm tắt đề án tốt nghiệp.
4. Phản biện 1 đọc nhận xét (có văn bản kèm theo)
5. Phản biện 2 đọc nhận xét (có văn bản kèm theo)
6. Các câu hỏi của thành viên Hội đồng:

a/ Tại sao chọn K=200, việc lựa chọn dựa trên tiêu chí nào?
b/ Việc áp dụng các mô hình xử lý mất cân bằng dữ liệu dựa trên SMOTE sẽ ảnh hưởng như thế nào?
c/ Nếu giao dịch nhiễu thì sẽ thực hiện như thế nào?

7. Trả lời của học viên:

a/ Tham số k hiện tại xác định thủ công sẽ thử chỉnh theo thực tế

b/ Việc xử lý mất cân bằng dữ liệu sẽ giúp giải quyết vấn đề thời gian của dữ liệu giao dịch tài chính

c/ Học viên sẽ tiếp thu và nghiên cứu thêm

8. Thư ký đọc nhận xét về quá trình thực hiện đề án tốt nghiệp của học viên *(có văn bản kèm theo).*

9. Hội đồng họp riêng:

- Bầu Ban kiểm phiếu:
  1. Trưởng Ban kiểm phiếu: Đào Ngọc Phong
  2. Uỷ viên Ban kiểm phiếu: Nguyễn Thanh Tuyên
  3. Uỷ viên Ban kiểm phiếu: Đỗ Thanh Hà
- Hội đồng chấm đề án tốt nghiệp bằng bỏ phiếu kín.
- Ban kiểm phiếu làm việc:
- Trưởng Ban kiểm phiếu báo cáo kết quả kiểm phiếu *(có Biên bản họp Ban kiểm phiếu kèm theo)*
- **Điểm trung bình của đề án tốt nghiệp:** 8.6

Kết luận:

1. Các nội dung cần chỉnh sửa, hoàn thiện sau bảo vệ đề án tốt nghiệp:

a/ Rà soát và làm rõ, chỉnh sửa về
- Tập dữ liệu, việc xử lý mất cân bằng
- Mô hình và các thành phần triển khai
- Tiền chế, phương pháp và kết quả đánh giá

c/ Tiếp thu đầy đủ ý kiến của Hội đồng và 02 Phản biện

c/ Rà soát chỉnh sửa về trình bày
- Hình vẽ, sơ đồ
- Lời chỉnh tả, lỗi trình bày

2. Đề nghị Học viện công nhận (hoặc không) và cấp bằng (hoặc không) thạc sĩ cho học viên: Công nhận và cấp bằng

3. Đề án tốt nghiệp có thể phát triển thành đề tài nghiên cứu cho NCS.....Không.....

Buổi làm việc kết thúc vào..11h45... cùng ngày.

Chủ tịch	**Thư ký**
GS.TS. Từ Minh Phương	**TS. Đào Ngọc Phong**

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM

Độc lập – Tự do – Hạnh phúc

BẢN NHẬN XÉT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

(Dùng cho người phản biện)

Tên đề tài đề án tốt nghiệp: Ứng dụng Graph Neural Network và kỹ thuật xử lý dữ liệu mất cân bằng trong phát hiện giao dịch gian lận tài chính

Chuyên ngành: Khoa học máy tính.....

Mã chuyên ngành: 8.48.01.01.....

Họ và tên học viên: Trần Thị Hương Giang.....

Họ và tên người nhận xét: Đỗ Thanh Hà .....

Học hàm, học vị: Tiến sĩ .....

Chuyên ngành: Tin học .....

Cơ quan công tác: Khoa trí tuệ nhân tạo, Học viện công nghệ bưu chính viễn thông ...

Số điện thoại: 0913378868.....E-mail: dothanhha@ptit.edu.vn .....

NỘI DUNG NHẬN XÉT

I/ Cơ sở khoa học và thực tiễn, tính cấp thiết của đề tài:

Phát hiện gian lận là bài toán sống còn của các tổ chức tài chính. Điểm yếu của các phương pháp truyền thống là chỉ nhìn vào từng giao dịch đơn lẻ (tabular data). Đề án đã xác định đúng xu hướng hiện nay là khai thác **mối quan hệ mạng lưới** (graph-based) để phát hiện các hành vi gian lận tinh vi theo chuỗi hoặc nhóm tài khoản.

Đề án giải quyết trực diện hai thách thức lớn nhất của dữ liệu tài chính: (1) Cấu trúc quan hệ phức tạp và (2) Tỷ lệ mất cân bằng dữ liệu cực kỳ cao (dưới 0.1%).

II/ Nội dung của đề án tốt nghiệp, các kết quả đã đạt được:

Đề án xây dựng một Pipeline rất bài bản và có chiều sâu kỹ thuật:

- Gom cụm tài khoản: Đây là một bước xử lý thông minh. Việc dùng K-Means (k=200) để giảm độ thưa của đồ thị giúp mô hình GNN tập trung vào các "nhóm hành vi" thay vì các tài khoản đơn lẻ có tương tác quá ít.
- Mô hình GAT (Graph Attention Network): Việc chọn cơ chế *Attention* là phù hợp vì trong mạng lưới giao dịch, không phải mọi kết nối đều có mức độ rủi ro như nhau.

- SMOTE trên không gian Embedding: Đây là điểm sáng tạo nhất của đồ án. Thay vì tăng mẫu (oversampling) trên dữ liệu thô, học viên thực hiện trên không gian vector đã được GNN trích xuất đặc trưng. Điều này giúp các mẫu gian lận tổng hợp mang tính "ngữ cảnh" hơn.

Đề án sử dụng bộ dữ liệu PaySim (6.3 triệu bản ghi) – một tập dữ liệu tiêu chuẩn và đủ lớn để kiểm chứng các mô hình học sâu.

Kết quả đạt được: Kết quả thực nghiệm cho thấy sự khác biệt rõ rệt giữa hai nhánh (Baseline và SMOTE). Chỉ số PR-AUC tăng từ <0.03 lên 0.167 và F1-score từ 0 lên 0.178. Mặc dù con số tuyệt đối chưa cao, nhưng trong bối cảnh gian lận cực hiếm (<0.1%), đây là một bước tiến so với việc mô hình không thể học được gì (Baseline). Các kết quả cũng đc đối chiếu và so sánh với các công trình nghiên cứu mới.

III/ Những vấn đề cần giải thích thêm:

1. Việc thay đổi k có thể ảnh hưởng lớn đến mật độ đồ thị và hiệu năng của GAT, tuy nhiên tham số này được chọn cố định k=200 nhưng chưa giải trình rõ dựa trên cơ sở khoa học nào (ví dụ: phương pháp Elbow hay Silhouette).
2. SMOTE sinh mẫu mới bằng cách nội suy tuyến tính giữa các điểm dữ liệu: Khi áp dụng SMOTE trên không gian Embedding (sau tầng GAT), liệu các điểm dữ liệu "tổng hợp" này có thực sự đại diện cho một hành vi gian lận hợp lý về mặt logic nghiệp vụ tài chính, hay chỉ là các điểm toán học nằm giữa các vector?
3. Chỉ số PR-AUC, dù đã cải thiện nhưng mức 0.167 vẫn cho thấy tỷ lệ báo động giả (False Positive) còn khá cao (237 trường hợp trên 76 trường hợp đúng). Điều này có thể gây áp lực cho đội ngũ điều tra viên nếu triển khai thực tế.

4. Đề án triển khai trên Google Colab với GPU, nhưng chưa phân tích về độ trễ khi hệ thống phải xử lý dòng giao dịch phát sinh liên tục trong thực tế (Real-time inference).

IV/ Kết luận:

Đồ án của học viên Trần Thị Hương Giang thể hiện năng lực chuyên môn tốt trong lĩnh vực Khoa học máy tính và Học sâu. Đồng ý cho phép học viên bảo vệ đề án tốt nghiệp.

Ngày.....tháng...1....năm2026

NGƯỜI NHẬN XÉT

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

BẢN NHẬN XÉT ĐỀ ÁN THẠC SĨ

(Dùng cho người phản biện)

- Đề tài: **Ứng dụng GNN và kỹ thuật xử lý dữ liệu mất cân bằng trong phát hiện giao dịch gian lận tài chính.**
- Tác giả: Trần Thị Hương Giang.
- Chuyên ngành: Khoa học máy tính
- Mã số 8.48.01.01
- Người nhận xét: Nguyễn Linh Giang, TS, PGS. Công nghệ thông tin
- Cơ quan công tác: Trường CNTT&TT, ĐHBK HN

Nhận xét:

1. Tổng quan chung và tính cấp thiết

Đề án của học viên trình bày những nghiên cứu, tìm hiểu về bài toán phát hiện gian lận trong giao dịch tài chính, những phương pháp phát hiện gian lận, tập trung vào nhóm các phương pháp dựa trên học sâu, nhóm phương pháp sử dụng mạng Nơ-ron đồ thị. Học viên đã tìm hiểu các giai đoạn tiền xử lý dữ liệu, xây dựng mô hình biểu diễn dữ liệu tài chính bằng đồ thị, xử lý mất cân bằng dữ liệu. Học viên cũng đưa ra các tiêu chí đánh giá phương pháp theo các độ đo Precision-recall, ROC AUC để thể hiện năng lực phát hiện của mô hình.

Vấn đề phát hiện gian lận giao dịch tài chính có nhiều ứng dụng trong thực tế. Việc đưa thêm biểu diễn dạng đồ thị kết hợp với các phương pháp xử lý mất cân bằng dữ liệu làm cho phương pháp đề xuất hiệu quả hơn. Những vấn đề này có tính thực tiễn.

2. Ưu điểm, nhược điểm của luận văn về nội dung, hình thức

- Nội dung luận văn có phù hợp với đề tài và đáp ứng yêu cầu?

Luận văn trình bày về xây dựng và triển khai thử nghiệm phương pháp phát hiện gian lận giao dịch tài chính dựa trên biểu diễn mạng đồ thị và Mạng nơ-ron đồ thị. Trong quá trình xử lý dữ liệu có áp dụng một số phương pháp xử lý mất cân bằng dữ liệu. Luận văn của học viên Trần Thị Hương Giang tập trung vào những nội dung chính như sau:

+ Giới thiệu những tìm hiểu về bài toán phát hiện gian lận trong giao dịch tài chính, các cách tiếp cận. Học viên tìm hiểu những tiêu chí đánh giá, các yêu cầu, các mô hình ngữ cảnh.

+ Tìm hiểu về các phương pháp phân tích, biểu diễn dữ liệu đồ thị, tiền xử lý dữ liệu, phương pháp xử lý vấn đề mất cân bằng dữ liệu, và các phương pháp xây dựng mô hình, huấn luyện và thử nghiệm mô hình.

+ Triển khai thử nghiệm và đánh giá kết quả

Các nội dung này phù hợp với tên của đề tài đề án và đáp ứng yêu cầu đề án thạc sỹ kỹ thuật chuyên ngành Khoa học máy tính.

- Nhận xét chung:

+ Luận văn viết còn một số điểm về nội dung có thể hoàn thiện thêm. Những kết quả thử nghiệm và xử lý dữ liệu cần mô tả rõ hơn trong ngữ cảnh bài toán phát hiện gian lận trong giao dịch tài chính. Những kết quả thu được cho thấy khả năng nắm vấn đề của tác giả trong thực hiện luận văn.

+ Luận văn có một số lưu ý: cần nêu rõ các tiêu chí đánh giá thử nghiệm trong phát hiện gian lận.

3. Kết luận:

Tôi đồng ý để tác giả Trần Thị Hương Giang được bảo vệ đề án trước Hội đồng chấm đề án thạc sĩ Công nghệ thông tin, chuyên ngành Khoa học máy tính.

Câu hỏi:

- Phân tích rõ thêm các tiêu chí xây dựng và đánh giá phương pháp phát hiện gian lận trogn giao dịch tài chính.

Hà Nội, Ngày tháng 01 năm 2026

Người nhận xét



Nguyễn Linh Giang