ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

BÁO CÁO SEMINAR KHOA HỌC

FIFAR DATASET: A FRAUD DETECTION DATASET FOR LEARNING TO DEFER

Sinh viên: Đàm Văn Hiển - 22022664
Đào Duy Hưng - 22022589
Phạm Thành Long - 22022604
Hoàng Ngọc Hào - 22022668

Giảng viên đánh giá: PGS.TS Nguyễn Phương Thái
TS. Trần Hồng Việt

HÀ NỘI - 6/2025

LỜI MỞ ĐẦU

Chúng em xin chân thành cảm ơn thầy cô đã đóng góp ý kiến trong suốt quá trình thực hiện bài viết này. Những nhận xét và góp ý của thầy cô đã giúp chúng em hoàn thiện và nâng cao chất lượng bài viết rất nhiều.

Tuy nhiên, bài viết vẫn còn thiếu sót và sẽ tiếp tục hoàn thiện trong thời gian tới để có thể đáp ứng tốt hơn các yêu cầu về nội dung và chất lượng. Chúng em rất mong nhận được sự hỗ trợ thêm từ thầy cô để cải thiện bài viết này.

Mục lục

1 Đặt vấn đề

1.1 Bối cảnh: Thách thức trong phát hiện gian lận tài chính

Gian lận tài chính là một trong những vấn đề nghiêm trọng đối với các tổ chức tài chính và khách hàng, gây thiệt hại lớn cả về kinh tế lẫn uy tín. Các hệ thống phát hiện gian lận truyền thống dựa vào quy tắc hoặc giám sát thủ công không còn đủ hiệu quả trước quy mô và mức độ tinh vi ngày càng gia tăng của các hành vi gian lận. Trong những năm gần đây, trí tuệ nhân tạo (AI) đã chứng minh vai trò quan trọng trong việc tự động hóa quá trình phát hiện các giao dịch bất thường, giúp tiết kiệm thời gian và nâng cao độ chính xác.

Tuy nhiên, AI vẫn còn nhiều hạn chế: (i) khó diễn giải quyết định, (ii) dễ bị sai lệch nếu dữ liệu huấn luyện chứa thiên kiến, và (iii) thiếu khả năng thích ứng với các tình huống mới hoặc không phổ biến. Điều này dẫn đến sự cần thiết phải phối hợp giữa AI và con người, trong đó AI xử lý các trường hợp phổ biến, còn con người xử lý các tình huống khó hoặc có yếu tố bất định cao.

1.2 Hợp tác con người - AI và vai trò của học cách ủy quyền (Learning to Defer)

Để tối ưu hóa sự phối hợp giữa AI và con người, lĩnh vực học máy đã phát triển hướng nghiên cứu gọi là *Learning to Defer* (L2D) — học cách ủy quyền quyết định. Thay vì để AI tự động đưa ra tất cả các quyết định, L2D cho phép hệ thống đánh giá mức độ tự tin trong dự đoán của mình và lựa chọn "ủy quyền" quyết định sang con người trong những tình huống không chắc chắn. Cách tiếp cận này giúp tận dụng tối đa năng lực của cả AI và con người, đồng thời giảm thiểu các sai sót nghiêm trọng.

Tuy nhiên, nhiều nghiên cứu hiện tại về L2D giả định rằng con người luôn sẵn sàng và có khả năng xử lý tất cả các yêu cầu được chuyển giao. Trên thực tế, các chuyên gia chỉ có năng lực xử lý giới hạn — do giới hạn thời gian, tài nguyên và khối lượng công việc — điều này làm phát sinh nhu cầu mô hình hóa các ràng buộc năng suất của con người trong nghiên cứu L2D.

1.3 Khoảng trống nghiên cứu: Thiếu dữ liệu thực tế hỗ trợ L2D

Một rào cản lớn trong việc phát triển và đánh giá các hệ thống L2D chính là sự thiếu hụt các tập dữ liệu công khai có mô phỏng thực tế sự hợp tác AI - con người. Đặc biệt, các tập dữ liệu hiện có thường không phản ánh đúng những yếu tố quan trọng như:

- Sự đa dạng trong cách ra quyết định của con người (thiên kiến cá nhân, kinh nghiệm chuyên môn).
- Giới hạn về năng suất xử lý của chuyên gia.
- Mức độ bất định trong dự đoán và đánh giá giữa AI và con người.

Điều này dẫn đến việc nhiều mô hình L2D được phát triển trong môi trường giả lập, thiếu tính thực tiễn và khó áp dụng hiệu quả trong các hệ thống thật.

1.4 Động lực phát triển FiFAR Dataset

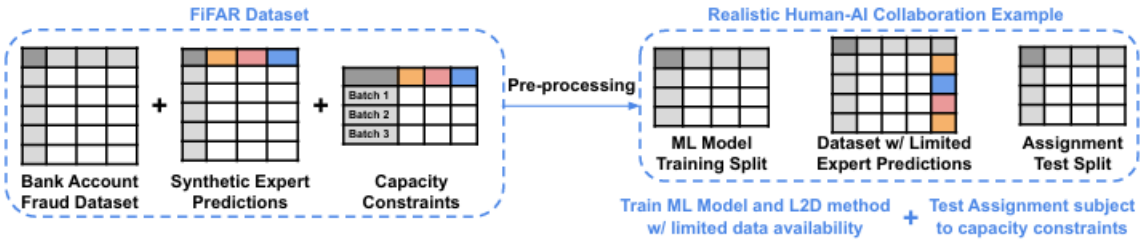
Nhằm khắc phục các hạn chế nêu trên, nhóm tác giả giới thiệu tập dữ liệu tổng hợp **FiFAR (Financial Fraud Alert Review)**. Đây là bộ dữ liệu đầu tiên mô phỏng chi tiết các quyết định của 50 chuyên gia phân tích gian lận với đặc điểm khác biệt về phong cách đánh giá, thiên kiến và giới hạn năng suất. FiFAR được xây dựng từ ba thành phần chính:

1. Dữ liệu thực tế về các giao dịch tài chính đã được gán nhãn.
2. Các quyết định giả lập từ chuyên gia tổng hợp (synthetic experts).
3. Ràng buộc về năng lực xử lý của từng chuyên gia trong môi trường làm việc thực tế.

FiFAR tạo điều kiện lý tưởng để kiểm thử các thuật toán L2D trong hơn 300 kịch bản khác nhau, từ đó đánh giá cả độ chính xác và tính công bằng của mô hình. Bộ dữ liệu này mở ra hướng nghiên cứu mới về việc tối ưu hóa sự phối hợp giữa con người và AI, đặc biệt trong các lĩnh vực nhạy cảm như tài chính — nơi mà các sai sót trong quyết định có thể dẫn đến hậu quả nghiêm trọng.

2 Cơ sở lí thuyết

2.1 FIFAR Dataset



Hình 1: Chia tập dữ liệu theo thời gian để phát triển L2D

2.1.1 Giới thiệu về tập dữ liệu FiFAR

Tập dữ liệu Financial Fraud Alert Review (FiFAR) là một bộ dữ liệu tổng hợp quy mô lớn, được xây dựng từ ba yếu tố chính: dữ liệu thực tế về gian lận tài chính, dự đoán từ các chuyên gia mô phỏng, và các giới hạn về năng lực xử lý. Thành phần đầu tiên, Bank Account Fraud Dataset, bao gồm thông tin về các giao dịch tài chính đã được phân loại là gian lận hoặc hợp pháp, tạo cơ sở thực tiễn cho việc phát triển các mô hình phát hiện gian lận. Thành phần thứ hai, Synthetic Expert Predictions, bao gồm các dự đoán được tạo ra bởi 50 chuyên gia mô phỏng, tái hiện quá trình ra quyết định của con người với các đặc tính riêng, như thiên vị nhận thức và phong cách đánh giá khác biệt. Mỗi chuyên gia có cách tiếp cận riêng, dựa trên các đặc điểm giao dịch như giá trị, địa điểm, hoặc lịch sử tài khoản. Thành phần cuối, Capacity Constraints, phản ánh thực tế về giới hạn năng lực con người, quy định số lượng giao dịch mà mỗi chuyên gia có thể xử lý trong một khoảng thời gian nhất định.

FiFAR bao gồm hơn một triệu giao dịch có nhãn, trở thành một trong những bộ dữ liệu lớn nhất hỗ trợ nghiên cứu về thuật toán Learning to Defer (L2D). Khác với các bộ dữ liệu truyền thống về gian lận, FiFAR tái tạo môi trường hợp tác giữa AI và con người bằng cách tích hợp thông tin về hạn chế trong khả năng ra quyết định của con người. Trong FiFAR, mỗi chuyên gia có độ chính xác và khuynh hướng đánh giá riêng, chẳng hạn như xu hướng gắn nhãn quá nhiều giao dịch là gian lận (tỷ lệ dương tính giả cao) hoặc bỏ qua nhiều giao dịch gian lận thực sự (tỷ lệ âm tính giả cao). Do

giới hạn về thời gian và khối lượng công việc, các chuyên gia không thể xử lý tất cả giao dịch, buộc AI phải học cách phân bổ nhiệm vụ một cách hợp lý.

Bộ dữ liệu này nhằm thu hẹp khoảng cách giữa lý thuyết L2D và các tình huống thực tế trong hợp tác giữa con người và AI. Sự kết hợp giữa dữ liệu thực, dự đoán của chuyên gia mô phỏng, và các giới hạn xử lý tạo ra một môi trường lý tưởng để đánh giá các thuật toán trì hoãn. Nhờ đó, FiFAR không chỉ kiểm tra hiệu suất của mô hình AI mà còn hỗ trợ nghiên cứu cách tối ưu hóa sự hợp tác giữa AI và con người trong phát hiện gian lận tài chính.

2.1.2 Phương pháp tạo chuyên gia tổng hợp

Một điểm nổi bật của FiFAR là khung tạo chuyên gia mô phỏng, tái hiện cách các nhà phân tích gian lận đưa ra quyết định. Khung này đảm bảo rằng các chuyên gia không chỉ được gán nhãn ngẫu nhiên mà tuân theo quy trình ra quyết định có cấu trúc, chịu ảnh hưởng từ nhiều yếu tố.

Dự đoán của mỗi chuyên gia được tạo dựa trên mô hình nhiễu phụ thuộc vào đặc điểm giao dịch (instance-dependent noise). Cụ thể, xác suất mắc lỗi của từng chuyên gia được điều chỉnh bởi các tham số phản ánh mức độ ảnh hưởng của đặc điểm giao dịch và điểm số từ mô hình phát hiện gian lận.

Xác suất lỗi của chuyên gia được xác định bởi các công thức:

$$\mathbb{P}(m_{i,j} = 1 | y_i = 0, \mathbf{x}_i, M) = \sigma\left(\beta_0 - \alpha \frac{\mathbf{w} \cdot \mathbf{x}_i + w_M M(\mathbf{x}_i)}{\sqrt{\mathbf{w} \cdot \mathbf{w} + w_M^2}}\right) \tag{1}$$

$$\mathbb{P}(m_{i,j} = 0 | y_i = 1, \mathbf{x}_i, M) = \sigma\left(\beta_1 + \alpha \frac{\mathbf{w} \cdot \mathbf{x}_i + w_M M(\mathbf{x}_i)}{\sqrt{\mathbf{w} \cdot \mathbf{w} + w_M^2}}\right) \tag{2}$$

trong đó $M(\mathbf{x}_i)$ là điểm số chuẩn hóa từ mô hình phát hiện gian lận, giúp chuyên gia phản ứng dựa trên độ tin cậy của mô hình.

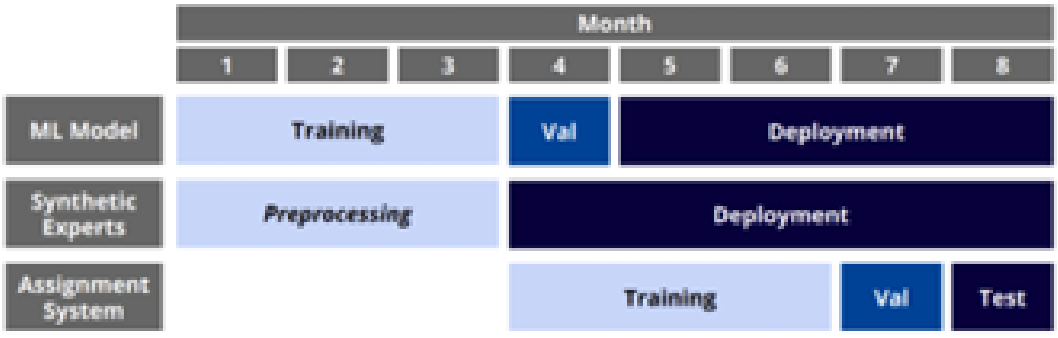
Phương pháp này cho phép FiFAR tạo ra các chuyên gia với thiên lệch có hệ thống, như Chuyên gia thận trọng (đặt ngưỡng cao để xác định gian lận, dẫn đến tỷ lệ âm tính giả cao), Chuyên gia nhạy bén (gắn nhãn quá nhiều giao dịch là gian lận, gây ra tỷ lệ dương tính giả cao), hoặc Chuyên gia tập trung đặc điểm (chỉ chú ý đến

một số yếu tố như giá trị giao dịch hoặc vị trí địa lý, khiến quyết định bị thiên lệch).

Cách tiếp cận này giúp FiFAR tái hiện sự đa dạng trong quyết định của các chuyên gia và tạo ra bộ dữ liệu phản ánh tính phức tạp của phát hiện gian lận thực tế. Điều này hỗ trợ thử nghiệm các thuật toán Learning to Defer (L2D) trong môi trường thực tiễn, nơi AI cần phân bổ quyết định hiệu quả giữa mô hình và con người.

2.1.3 Phương pháp L2D và đánh giá

Để đánh giá các chiến lược Learning to Defer (L2D), các nhà nghiên cứu so sánh nhiều mô hình ra quyết định. Rejection Learning (ReL) học cách từ chối các trường hợp phức tạp và chuyển chúng cho chuyên gia con người, giảm thiểu lỗi phân loại. Greedy Assignment sử dụng cách tiếp cận heuristic đơn giản, chuyển giao dịch cho con người dựa trên ngưỡng độ tin cậy cố định. Integer Linear Programming (ILP) là phương pháp tối ưu hóa, phân bổ giao dịch cho AI hoặc con người để tối ưu hiệu suất, tuân thủ giới hạn năng lực của chuyên gia. Hiệu suất được đo bằng hai chỉ số: Recall tại mức 5% False Positive Rate (FPR), đảm bảo phát hiện gian lận hiệu quả với tỷ lệ báo động giả thấp, và Predictive Equality, đánh giá sự công bằng trong phân bố lỗi giữa các nhóm giao dịch. Các tiêu chí này đảm bảo FiFAR không chỉ đánh giá độ chính xác mà còn kiểm tra tính công bằng và khả năng áp dụng thực tế.



Hình 2: Chia tập dữ liệu theo thời gian để phát triển L2D

2.1.4 Những phát hiện và đóng góp chính

Kết quả nghiên cứu làm sáng tỏ các hiểu biết quan trọng về phát hiện gian lận và chiến lược Learning to Defer. Các mô hình L2D cải thiện đáng kể hiệu quả phát hiện gian lận, vượt trội hơn so với các mô hình chỉ dựa vào AI hoặc con người. Hiệu

suất giảm khi chuyên gia con người bị quá tải, nhấn mạnh tầm quan trọng của L2D trong việc xem xét giới hạn năng lực con người. Phân bổ giao dịch thông minh giúp giảm chênh lệch lỗi giữa các nhóm giao dịch. FiFAR là bộ dữ liệu lớn đầu tiên mô phỏng các giới hạn ra quyết định của con người trong phát hiện gian lận, cung cấp tiêu chuẩn để đánh giá thuật toán L2D qua 300 kịch bản thử nghiệm và giới thiệu các mô hình L2D mới, tối ưu hóa hợp tác giữa AI và con người trong điều kiện thực tế. Những đóng góp này khiến FiFAR trở thành tài nguyên quan trọng, thúc đẩy nghiên cứu về phát hiện gian lận và ra quyết định hỗ trợ bởi AI.

2.2 Các vấn đề liên quan với bài toán HAIC

Trong bối cảnh ngày càng phổ biến của các hệ thống hỗ trợ ra quyết định, sự hợp tác giữa con người và AI (Human-AI Collaboration — HAIC) không chỉ là một vấn đề kỹ thuật, mà còn đặt ra hàng loạt câu hỏi về đạo đức, công bằng, khả năng diễn giải và tính khả thi triển khai. Bài toán phát hiện gian lận tài chính là một ví dụ điển hình cho các thách thức này, nơi mà quyết định sai có thể gây tổn hại nghiêm trọng đến quyền lợi cá nhân và niềm tin xã hội. Dưới đây là bốn vấn đề trọng yếu cần được xem xét khi phát triển và đánh giá các hệ thống HAIC như trong FiFAR.

2.2.1 Đạo đức AI trong các quyết định quan trọng

Trong các lĩnh vực có rủi ro cao như tài chính, đạo đức AI trở thành yêu cầu bắt buộc. Quyết định từ chối mở tài khoản ngân hàng hay từ chối giao dịch do nghi ngờ gian lận có thể ảnh hưởng đến khả năng tiếp cận dịch vụ tài chính của một cá nhân. Nếu hệ thống AI hoạt động như một "hộp đen" và không có cơ chế chịu trách nhiệm hoặc khiếu nại, nó có thể xâm phạm quyền cá nhân và gây ra bất công.

FiFAR giải quyết vấn đề này bằng cách mô phỏng các tình huống ra quyết định có yếu tố đạo đức, nơi AI cần biết khi nào nên tự ra quyết định và khi nào nên ủy quyền cho con người. Việc này đặt ra câu hỏi: liệu hệ thống có đưa ra quyết định dựa trên độ tin cậy hay dựa trên rủi ro gây tổn hại cho người dùng? Từ đó, các mô hình L2D phải được thiết kế để đảm bảo tính minh bạch và khả năng can thiệp của con người trong các tình huống rủi ro cao.

2.2.2 Sự công bằng và thiên lệch trong các hệ thống AI

Một trong những rủi ro lớn nhất của AI là khả năng tái tạo và khuếch đại các thiên lệch có trong dữ liệu huấn luyện. Trong phát hiện gian lận tài chính, các mô hình có thể gắn nhãn sai các giao dịch hợp pháp từ người cao tuổi là "gian lận" nhiều hơn so với các nhóm khác. Điều này dẫn đến nguy cơ loại trừ tài chính có hệ thống — một biểu hiện rõ ràng của bất công.

FiFAR cho phép đánh giá tính công bằng qua chỉ số *Predictive Equality (PE)* — tỉ lệ chênh lệch trong False Positive Rate giữa các nhóm dân số. Hệ thống L2D được kiểm tra trên hơn 300 kịch bản để đảm bảo rằng việc ủy quyền không gây thiên lệch mới hoặc khuếch đại thiên lệch sẵn có. Đặc biệt, FiFAR còn mô phỏng các chuyên gia tổng hợp với mức độ thiên lệch khác nhau, từ đó kiểm tra khả năng của hệ thống trong việc tránh gán quyết định cho các chuyên gia thiếu công bằng.

2.2.3 Khả năng diễn giải và tính minh bạch của mô hình

Một yêu cầu thiết yếu của HAIC là mô hình AI phải có khả năng giải thích được quyết định của mình, đặc biệt khi người dùng cuối là các chuyên gia cần phối hợp để xử lý các trường hợp phức tạp. Trong hệ thống L2D, không chỉ cần giải thích tại sao AI đưa ra dự đoán mà còn phải giải thích tại sao chọn giao cho con người xử lý thay vì tự quyết định.

FiFAR thúc đẩy việc tích hợp các phương pháp như SHAP, LIME để cải thiện diễn giải mô hình. Trong quá trình huấn luyện và đánh giá, các thuật toán cần cung cấp lý do rõ ràng khi chuyển giao quyết định — chẳng hạn như độ không chắc chắn cao, độ tương đồng thấp với các mẫu trong tập huấn luyện, hoặc rủi ro cao của false positive. Việc này giúp tăng độ tin cậy của người dùng vào hệ thống, đồng thời tạo điều kiện để thiết lập cơ chế giám sát và phản hồi phù hợp.

2.2.4 Những thách thức triển khai mô hình trong thực tế

Dù các mô hình L2D mang lại nhiều hứa hẹn về mặt lý thuyết, triển khai chúng trong thực tế đòi hỏi giải quyết nhiều yếu tố phức tạp. Một vấn đề nổi bật là giới hạn năng lực của chuyên gia con người — họ không thể xử lý vô số trường hợp được AI ủy quyền. Thực tế này được FiFAR mô phỏng rõ ràng thông qua các "ràng buộc năng

suất" (capacity constraints), nơi mỗi chuyên gia chỉ có thể xử lý một số lượng giao dịch giới hạn trong mỗi lô (batch) dữ liệu.

Ngoài ra, môi trường thực tế rất năng động — hành vi gian lận thay đổi theo thời gian, quy định pháp lý liên tục cập nhật, và khối lượng dữ liệu ngày càng tăng. Việc huấn luyện và cập nhật mô hình thường xuyên với sự hỗ trợ của con người là khó khăn và tốn kém. FiFAR tạo điều kiện để kiểm tra hệ thống L2D trong các kịch bản có thay đổi về mặt nhân lực (chuyên gia vắng mặt), khối lượng công việc (batch size lớn), và tỷ lệ chuyển giao (deferral rate thay đổi), giúp mô phỏng và kiểm định tính khả thi trước khi triển khai vào môi trường thực tế.

3 Thực nghiệm và đánh giá

Mã nguồn: https://github.com/longluv1605/fifar-seminar

3.1 Thiết lập thí nghiệm

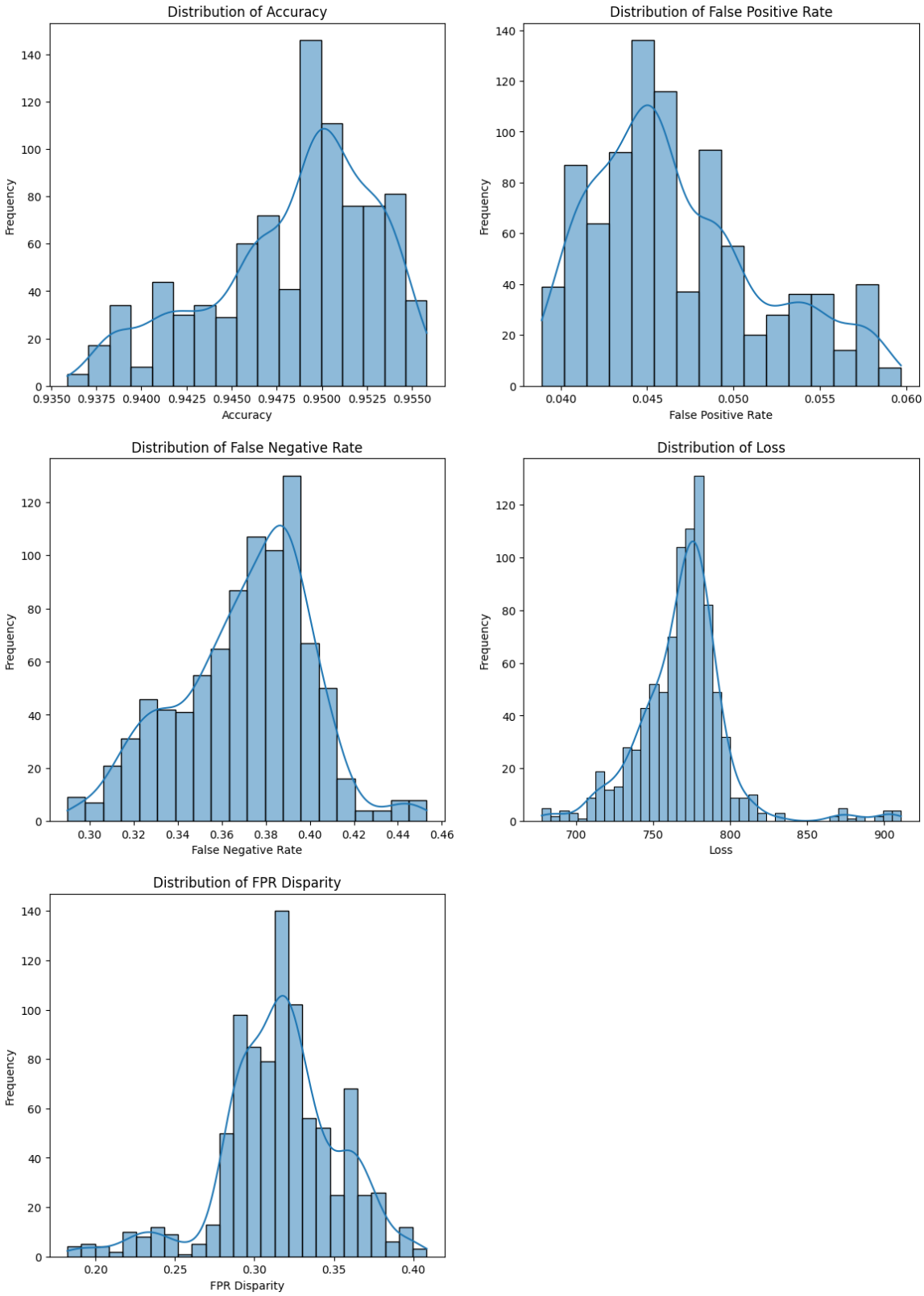
Trong phần này, nhóm em tiến hành thực nghiệm với bài toán Bank Account Fraud Detection. Trước hết, cần phải chuẩn bị dữ liệu, bao gồm: (1) **Bank Account Fraud Tabular Dataset**; (2) **Dữ liệu và mô hình cần thiết cho FIFAR** chứa mô hình chuyên gia, mô hình phân loại, dữ liệu chuyên gia và dữ liệu thử nghiệm.

Sau khi có được dữ liệu ban đầu, nhóm thực hiện phân bổ dữ liệu vào các module của thí nghiệm, đồng thời huấn luyện và hiệu chỉnh các mô hình.

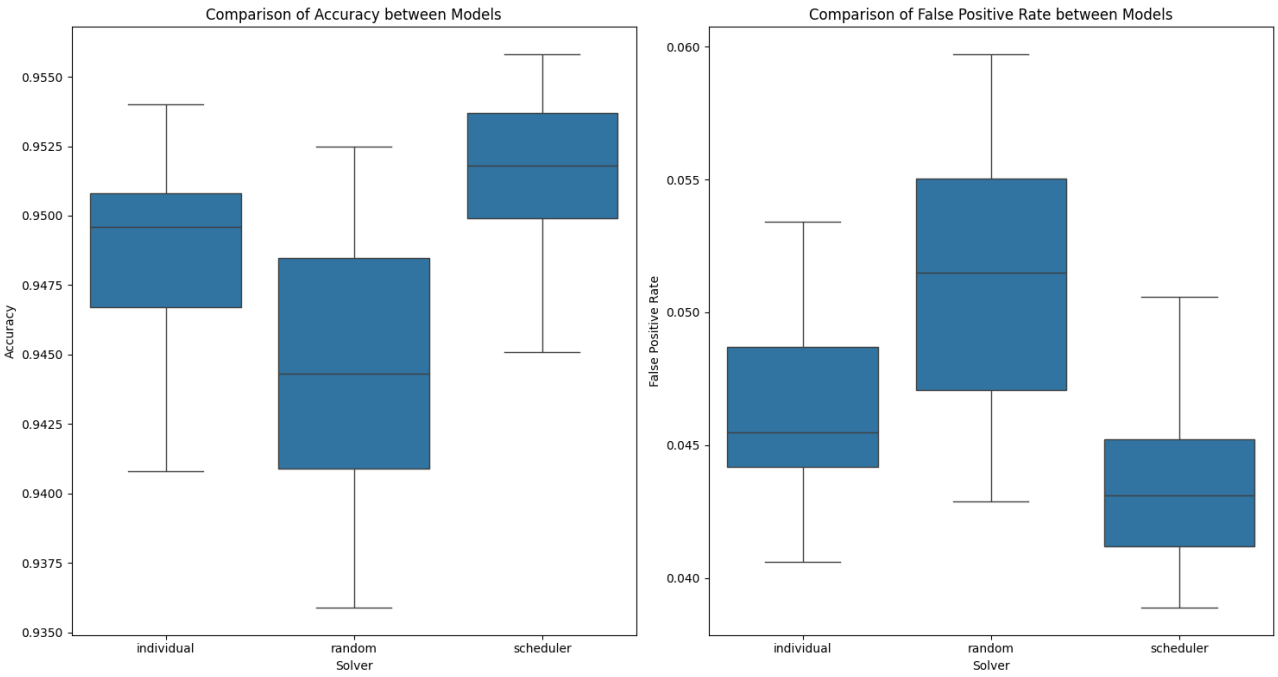
Dữ liệu thực nghiệm được thiết kế với 3 cài đặt của chiến lược phân công (assignment strategy): (1) *individual* - phân công độc lập từng trường hợp; (2) *random* - phân công ngẫu nhiên chuyên gia; (3) *scheduler* - chiến lược phân công tối ưu toàn cục.

Dữ liệu test gồm có **300** kịch bản với các thiết đặt khác nhau về batch và capacity, tổng cộng cần thực nghiệm trên **900** kịch bản. Quá trình phân công sử dụng framework *autodefer* do các tác giả của bài báo cung cấp. Quá trình thử nghiệm có thể được bắt đầu lại bất cứ khi nào mà không cần phải thực hiện lại các thử nghiệm đã làm.

3.2 Kết quả



Hình 3: Biểu đồ thể hiện phân bố của các chỉ số



Hình 4: Biểu đồ so sánh Accuracy và FPR giữa các chiến lược phân công

3.3 Đánh giá

Dựa trên quá trình thực nghiệm và các biểu đồ kết quả thu được, em rút ra một số nhận xét sau:

- **Phân phối các chỉ số đánh giá**: Các biểu đồ phân phối cho thấy độ chính xác (accuracy) của các thuật toán phân công chủ yếu tập trung ở mức cao, tuy nhiên vẫn tồn tại sự phân tán nhất định giữa các kịch bản. Tỷ lệ dương tính giả (FPR) và tỷ lệ âm tính giả (FNR) có sự biến động rõ rệt, phản ánh sự khác biệt về hiệu quả giữa các thuật toán và kịch bản thử nghiệm. Chỉ số chi phí (loss) và sự chênh lệch tỷ lệ dương tính giả (FPR disparity) cũng cho thấy một số trường hợp có giá trị cao, cho thấy vẫn còn tồn tại các kịch bản khó tối ưu.

- **So sánh giữa các thuật toán phân công (solver)**: Biểu đồ box plot cho thấy thuật toán *scheduler* thường đạt kết quả tốt nhất về độ chính xác và tỷ lệ dương tính giả, với phân phối ổn định và ít outlier. Trong khi đó, các thuật toán *individual* và *random* có độ phân tán lớn hơn, đặc biệt là *random* thường cho kết quả kém ổn định và hiệu quả thấp hơn rõ rệt. Điều này cho thấy việc sử dụng các thuật toán tối ưu hóa toàn cục như *scheduler* là cần thiết để đảm bảo hiệu quả và tính công bằng cho hệ thống.

Từ những nhận định trên, có thể thấy rằng việc áp dụng thuật toán phân công tối ưu toàn cục (*scheduler*) là lựa chọn phù hợp cho các hệ thống cần đảm bảo hiệu quả tổng thể và fairness. Các thuật toán baseline như *individual* và *random* chủ yếu phù hợp để so sánh đối chứng, nhưng không đáp ứng tốt các yêu cầu về hiệu quả và độ ổn định trong thực tế.

4 Kết luận

Học trì hoãn quyết định (Learning to Defer – L2D) là một hướng nghiên cứu có tiềm năng lớn trong việc xây dựng các hệ thống ra quyết định kết hợp giữa AI và con người, đặc biệt trong các lĩnh vực rủi ro cao như tài chính. Tuy nhiên, tiến bộ trong lĩnh vực này đã bị hạn chế đáng kể bởi sự thiếu hụt dữ liệu thực tế mô phỏng đúng hành vi và năng lực của con người.

Bộ dữ liệu FiFAR đã ra đời như một đóng góp đáng kể nhằm giải quyết vấn đề này. Với khả năng mô phỏng hành vi chuyên gia đa dạng và tích hợp giới hạn năng lực xử lý, FiFAR cho phép đánh giá và so sánh các thuật toán L2D trong bối cảnh thực tiễn sát với các hệ thống thật. Không chỉ dừng lại ở việc cung cấp dữ liệu, FiFAR còn đưa ra một khung thử nghiệm rõ ràng, có thể mở rộng, giúp đánh giá được cả hiệu suất lẫn tính công bằng của mô hình.

Việc triển khai thực nghiệm thông qua dự án đã góp phần làm rõ hơn vai trò và hiệu quả của các thuật toán phân công trong các hệ thống L2D. Kết quả phân tích cho thấy khả năng mô phỏng sát thực tế và kiểm soát được các yếu tố như độ lệch, giới hạn xử lý hay sự vắng mặt của chuyên gia là yếu tố then chốt để đảm bảo tính ổn định và hiệu quả khi đánh giá mô hình.

Nhìn chung, FiFAR không chỉ giúp khắc phục rào cản dữ liệu trong nghiên cứu L2D, mà còn mở ra một hướng đi mới cho việc thiết kế và thử nghiệm các hệ thống hỗ trợ quyết định công bằng, linh hoạt và có thể mở rộng.

Trong tương lai, dự án có thể mở rộng và phát triển theo các hướng sau:

- Nâng cấp bộ dữ liệu với số lượng chuyên gia tổng hợp đa dạng hơn, mô phỏng hành vi thực tế chi tiết hơn.
- Áp dụng và so sánh với các thuật toán L2D tiên tiến hơn như mô hình học sâu tích hợp phân công.

- Đánh giá sâu hơn về tính công bằng theo các tiêu chuẩn khác nhau (demographic parity, equalized odds...).
- Phân tích tác động của sai lệch thuật toán (algorithmic bias) và chiến lược giảm thiểu.

Tài liệu tham khảo

[1] J. V. Alves, D. Leitão, S. Jesus, M. O. P. Sampaio, P. Saleiro, M. A. T. Figueiredo, and P. Bizarro, "Fifar: A fraud detection dataset for learning to defer," 2023. [Online]. Available: https://arxiv.org/abs/2312.13218

[2] C.-C. Chen and M. Z. Yao, "Strategic use of immersive media and narrative message in virtual marketing: Understanding the roles of telepresence and transportation," *Psychology and Marketing*, vol. 39, no. 3, pp. 524–542, 2022. [Online]. Available: https://doi.org/10.1002/mar.21630

[3] C. Elkan, "The foundations of cost-sensitive learning," in *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI 2001)*. Morgan Kaufmann, 2001, pp. 973–978.

[4] L. Alves. (2023) Fifar - financial fraud alert review dataset. [Online]. Available: https://www.kaggle.com/datasets/leonardovalves/fifar-financial-fraud-alert-review-dataset/data

[5] sgpjesus. (2022) Bank account fraud dataset (neurips 2022). [Online]. Available: https://www.kaggle.com/datasets/sgpjesus/bank-account-fraud-dataset-neurips-2022

Nhận xét của Thầy/Cô

. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .

Hà Nội, ngày , tháng năm 2025

Ký tên

(Ký và ghi rõ họ tên)