FiFAR: A Fraud Detection Dataset for Learning to Defer

Phạm Thành Long
UET-VNU, Viet Nam
22022604@edu.vnu.vn
Đàm Văn Hiển
UET-VNU, Viet Nam
22022664@edu.vnu.vn
Đào Duy Hưng
UET-VNU, Viet Nam
22022589@edu.vnu.vn
Hoàng Ngọc Hào
UET-VNU, Viet Nam
22022668@edu.vnu.vn

Tóm tắt

Thiếu hụt tập dữ liệu công khai là một rào cản lớn trong nghiên cứu các thuật toán học cách ủy quyền (L2D), vốn nhằm tối ưu hóa sự phối hợp giữa AI và con người trong các hệ thống ra quyết định lai. Trong phát hiện gian lận tài chính, AI có thể nhanh chóng xác định các giao dịch đáng ngờ, nhưng vẫn cần chuyên gia con người để đánh giá bối cảnh và xử lý các trường hợp phức tạp. Tuy nhiên, hiện chưa có tập dữ liệu công khai nào hỗ trợ nghiên cứu L2D trong bối cảnh này.

Chúng tôi giới thiệu **FiFAR**, một tập dữ liệu tổng hợp mô phỏng quyết định của 50 chuyên gia phân tích gian lận với các thiên kiến và giới hạn năng suất khác nhau. FiFAR không chỉ hỗ trợ nghiên cứu L2D trong điều kiện thực tế mà còn cho phép đánh giá các chiến lược phân công nhiệm vụ một cách khách quan.

Chúng tôi thực hiện đánh giá trên 300 kịch bản kiểm thử để kiểm tra hiệu quả của các phương pháp L2D, từ đó đưa ra những phân tích quan trọng về tác động của giới hạn năng suất lên độ chính xác và tính công bằng trong phát hiện gian lận.

Tập dữ liệu và thông tin chi tiết có sẵn tại: https://github.com/feedzai/fifar-dataset.

Keywords: Learning to defer, Human-AI collaboration, Fraud detection, Deferral algorithms, FiFAR dataset, AI ethics, Fairness in AI.

I. Giới thiệu

Phát hiện gian lận tài chính đóng vai trò quan trọng trong việc bảo vệ hệ thống ngân hàng và khách hàng khỏi tổn thất nghiêm trọng. Trong lĩnh vực này, AI được sử dụng để phân tích lượng lớn giao dịch và phát hiện các mẫu bất thường, giúp xử lý gian lận ở quy mô lớn. Tuy nhiên, AI vẫn tồn tại nhiều hạn chế, bao gồm thiên kiến trong dữ liệu huấn luyện, khả năng bị khai thác bởi các chiến thuật gian lận mới và khó khăn trong việc diễn giải quyết định của mô hình.

Ngược lại, chuyên gia con người có thể tận dụng kinh nghiệm, trực giác và hiểu biết bối cảnh để xác định các hành vi gian lận tinh vi mà AI có thể bỏ sót. Sự phối hợp giữa AI và con người giúp cải thiện độ chính xác, giảm thiểu cảnh báo sai và nâng cao tính công bằng trong phát hiện gian lận. Tuy nhiên, một thách thức quan trọng là làm thế nào để xác định khi nào AI nên tự ra quyết định và khi nào nên chuyển giao nhiệm vụ cho con người nhằm tối ưu hiệu suất chung.

Learning to Defer (L2D) là phương pháp giúp hệ thống ra quyết định xác định khi nào AI có đủ độ tin cậy để tự đưa ra quyết định và khi nào cần sự can thiệp của con người. Mục tiêu của L2D là tối ưu hóa hiệu suất tổng thể bằng cách tận dụng thế mạnh của cả AI và con người.

Tuy nhiên, triển khai L2D trong thực tế gặp nhiều thách thức. Một số phương pháp giả định rằng con người luôn sẵn sàng hỗ trợ, nhưng trong môi trường làm việc thực tế, năng suất của chuyên gia là có giới hạn. Quá tải công việc có thể dẫn đến sai sót hoặc quyết định kém chất lượng. Ngoài ra, việc thu thập dữ liệu gán nhãn đầy đủ từ cả AI và con người là tốn kém và khó thực hiện, gây cản trở cho quá trình huấn luyện và đánh giá mô hình.

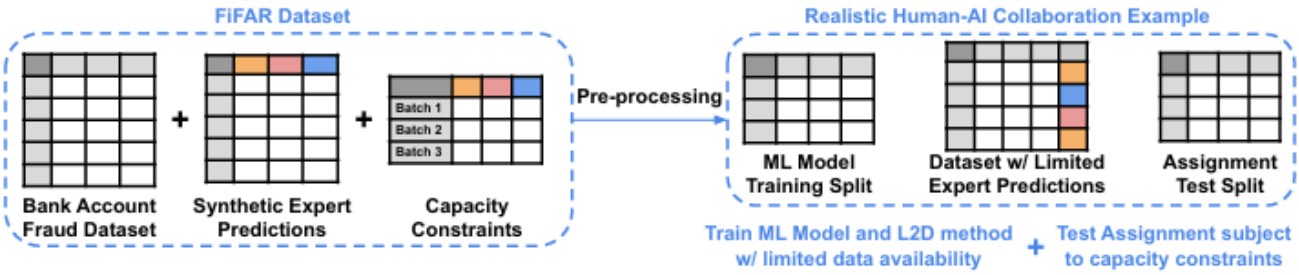
Nhằm giải quyết những hạn chế trên, chúng tôi giới thiệu **FiFAR**, một tập dữ liệu tổng hợp quy mô lớn với 1 triệu giao dịch được gán nhãn, mô phỏng quyết định của 50 chuyên gia phân tích gian lận có phong cách ra quyết định khác nhau. Không giống như các tập dữ liệu trước đây, FiFAR tích hợp yếu tố giới hạn năng suất của con người, giúp phản ánh chính xác hơn điều kiện thực tế trong các hệ thống L2D.

Dựa trên FiFAR, chúng tôi phân tích nhiều chiến lược L2D và đánh giá tác động của giới hạn năng suất lên hiệu suất phát hiện gian lận cũng như tính công bằng của hệ thống. Kết quả thực nghiệm trên 300 kịch bản kiểm thử cho thấy, việc xem xét yếu tố giới hạn năng suất giúp cải thiện đáng kể độ chính xác và tối ưu hóa sự phối hợp giữa AI và con người. FiFAR không chỉ là một nguồn dữ liệu quan trọng mà còn đặt nền tảng cho các nghiên cứu tiếp theo về L2D và hợp tác con người-AI.

II. Tổng quan về FiFAR

1. Giới thiệu về tập dữ liệu FiFAR

Tập dữ liệu **Financial Fraud Alert Review (FiFAR)** là một tập dữ liệu tổng hợp quy mô lớn được xây dựng từ ba thành phần chính: dữ liệu gian lận tài chính thực tế, dự đoán của các chuyên gia tổng hợp và các ràng buộc về khả năng xử lý. Thành phần đầu tiên, **Bank Account Fraud Dataset**, chứa thông tin về các giao dịch tài chính đã được gán nhãn gian lận hoặc hợp



Hình 1: FiFAR Dataset

lệ, cung cấp nền tảng thực tế để xây dựng mô hình phát hiện gian lận. Thứ hai, **Synthetic Expert Predictions** là tập hợp các dự đoán được tạo ra từ 50 chuyên gia tổng hợp, mô phỏng quá trình ra quyết định của con người với những đặc điểm riêng biệt, bao gồm thiên kiến nhận thức và xu hướng đánh giá khác nhau. Mỗi chuyên gia có phong cách phân tích riêng, dựa trên một số đặc điểm giao dịch nhất định, chẳng hạn như giá trị giao dịch, vị trí địa lý hoặc lịch sử hoạt động của tài khoản. Thành phần cuối cùng, **Capacity Constraints**, đóng vai trò quan trọng trong việc phản ánh thực tế về hạn chế của con người, giới hạn số lượng giao dịch mà mỗi chuyên gia có thể xử lý trong một khoảng thời gian nhất định.

Tập dữ liệu FiFAR bao gồm hơn một triệu giao dịch có nhãn, làm cho nó trở thành một trong những tập dữ liệu lớn nhất phục vụ nghiên cứu về thuật toán **Learning to Defer (L2D)**. Khác với các tập dữ liệu phát hiện gian lận truyền thống, FiFAR mô phỏng môi trường hợp tác giữa con người và AI bằng cách tích hợp thông tin về khả năng ra quyết định có giới hạn của con người. Trong FiFAR, mỗi chuyên gia không chỉ có độ chính xác khác nhau mà còn có những khuynh hướng đánh giá riêng, chẳng hạn như thiên hướng gắn nhãn quá nhiều giao dịch là gian lận (tỷ lệ dương tính giả cao) hoặc bỏ sót một số lượng lớn giao dịch gian lận thực sự (tỷ lệ âm tính giả cao). Ngoài ra, các chuyên gia này không thể đánh giá tất cả các giao dịch do giới hạn về thời gian và khối lượng công việc, khiến AI phải học cách phân bổ nhiệm vụ hợp lý.

Tập dữ liệu này hướng đến việc thu hẹp khoảng cách giữa các mô hình L2D lý thuyết và kịch bản hợp tác thực tế giữa con người và AI. Việc kết hợp dữ liệu thực tế, dự đoán của chuyên gia tổng hợp và ràng buộc về khả năng xử lý tạo ra một môi trường đánh giá lý tưởng cho các thuật toán trì hoãn. Bằng cách này, FiFAR không chỉ giúp kiểm tra hiệu quả của mô hình AI mà còn hỗ trợ nghiên cứu về cách AI có thể hợp tác với con người một cách tối ưu trong phát hiện gian lận tài chính.

2. *Phương pháp tạo chuyên gia tổng hợp*

Một trong những đổi mới quan trọng của FiFAR là khung tạo chuyên gia tổng hợp, mô phỏng cách các nhà phân tích gian lận thực sự đưa ra quyết định. Khung này đảm bảo rằng các chuyên gia không chỉ được gán nhãn ngẫu nhiên mà còn tuân theo các quy trình ra quyết định có cấu trúc, chịu ảnh hưởng bởi nhiều yếu tố khác nhau.

Dự đoán của mỗi chuyên gia không chỉ đơn thuần là một nhãn được gán ngẫu nhiên mà được tạo ra dựa trên mô hình *nhiễu phụ thuộc vào đặc điểm giao dịch* (instance-dependent noise). Cụ thể, mỗi chuyên gia có một xác suất mắc lỗi khác nhau, được điều chỉnh bởi các tham số kiểm soát mức độ ảnh hưởng của đặc điểm giao dịch và điểm số mô hình phát hiện gian lận.

Xác suất mắc lỗi của chuyên gia được xác định bởi 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 một phiên bản đã chuẩn hóa của điểm số mô hình phát hiện gian lận, đảm bảo rằng chuyên gia có thể phản ứng với sự tự tin của mô hình.

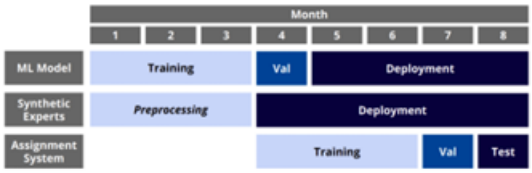
Dựa trên phương pháp này, FiFAR có thể tạo ra các chuyên gia với **các thiên kiến có hệ thống**, chẳng hạn như **Chuyên gia bảo thủ** (có ngưỡng rất cao để đánh dấu giao dịch là gian lận, dẫn đến **tỷ lệ âm tính giả cao**), **Chuyên gia nhạy cảm** (đánh dấu quá nhiều giao dịch là gian lận, tạo ra **tỷ lệ dương tính giả cao**) và **Chuyên gia theo đặc điểm cụ thể** (chỉ tập trung vào một số yếu tố nhất định của giao dịch, chẳng hạn như **giá trị giao dịch** hoặc **vị trí địa lý**, làm cho quyết định của họ bị thiên lệch theo các đặc điểm này).

Thông qua cách tiếp cận này, FiFAR không chỉ mô phỏng sự khác biệt trong quyết định của các chuyên gia mà còn tạo ra một tập dữ liệu phản ánh độ phức tạp của quá trình phát hiện gian lận trong thực tế. Điều này giúp thử nghiệm các thuật toán *Learning to Defer* (L2D) trong điều kiện thực tế, nơi mà hệ thống cần học cách phân bổ quyết định giữa mô hình và con người một cách hiệu quả.

3. *Phương pháp L2D và đánh giá*

Để đánh giá các chiến lược **Learning to Defer (L2D)**, các tác giả so sánh nhiều mô hình ra quyết định khác nhau. **Rejection Learning (ReL)** là một mô hình học cách từ chối các trường hợp khó và chuyển chúng đến

chuyên gia con người, giảm thiểu sai sót trong phân loại. **Greedy Assignment** là một phương pháp heuristic đơn giản, trong đó AI chuyển giao các giao dịch cho chuyên gia con người dựa trên ngưỡng độ tin cậy được xác định trước. **Integer Linear Programming (ILP)** là một phương pháp tối ưu hóa, gán giao dịch cho con người hoặc AI để tối ưu hóa hiệu suất trong khi vẫn tôn trọng các ràng buộc về khả năng xử lý của chuyên gia. Hiệu suất của các mô hình này được đo lường bằng hai chỉ số chính: **Recall tại mức 5% False Positive Rate (FPR)**, đảm bảo mô hình có thể phát hiện các trường hợp gian lận hiệu quả trong khi vẫn giữ tỷ lệ báo động giả ở mức thấp, và **Predictive Equality**, một chỉ số đo lường mức độ phân bố sai sót của mô hình phát hiện gian lận giữa các nhóm giao dịch khác nhau. Những tiêu chí đánh giá này đảm bảo rằng FiFAR không chỉ được sử dụng để đánh giá độ chính xác của mô hình mà còn để phân tích tính công bằng và khả năng áp dụng trong thực tế.



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

4. Những phát hiện và đóng góp chính

Các kết quả thực nghiệm của nghiên cứu này làm sáng tỏ một số hiểu biết quan trọng về phát hiện gian lận và chiến lược **Learning to Defer**. Các phương pháp L2D giúp cải thiện đáng kể hiệu suất phát hiện gian lận. Các mô hình chỉ sử dụng AI hoặc chỉ sử dụng con người hoạt động kém hiệu quả hơn so với các mô hình L2D kết hợp, vốn tận dụng sức mạnh của cả hai. Hiệu suất phát hiện gian lận giảm khi các chuyên gia con người bị quá tải, nhấn mạnh sự cần thiết của các mô hình L2D có tính đến năng suất của con người. Việc gán giao dịch một cách thông minh giúp giảm sự chênh lệch trong tỷ lệ lỗi giữa các loại giao dịch khác nhau. FiFAR là tập dữ liệu quy mô lớn đầu tiên mô hình hóa các ràng buộc ra quyết định của con người trong phát hiện gian lận. Nó cung cấp một tiêu chuẩn để đánh giá các thuật toán L2D trên 300 kịch bản thử nghiệm khác nhau và giới thiệu các mô hình L2D mới có khả năng tối ưu hóa sự hợp tác giữa con người và AI trong điều kiện thực tế. Những đóng góp này khiến FiFAR trở thành một nguồn tài nguyên quan trọng để thúc đẩy nghiên cứu trong lĩnh vực phát hiện gian lận và ra quyết định hỗ trợ bởi AI.

III. Các vấn đề liên quan với HAIC

1. Đạo đức AI với các quyết định quan trọng

Việc tích hợp trí tuệ nhân tạo (AI) vào các quy trình ra quyết định quan trọng đặt ra những thách thức đạo đức đáng kể, đặc biệt trong các ứng dụng tài chính, nơi quyền riêng tư, trách nhiệm giải trình và tính minh bạch đóng vai trò then chốt. Các hệ thống phát hiện gian lận do AI điều khiển, chẳng hạn, có vai trò quan trọng trong việc xác định liệu một cá nhân có thể tiếp cận dịch vụ ngân hàng, đăng ký vay vốn hoặc thực hiện các giao dịch tài chính hay không. Tuy nhiên, nếu không có các biện pháp bảo vệ phù hợp, các hệ thống này có thể vô tình hạn chế quyền tiếp cận các nguồn tài chính thiết yếu, gây ảnh hưởng không cân xứng đến một số nhóm dân cư nhất định.

Đảm bảo tính công bằng trong các ứng dụng này đòi hỏi sự cân bằng giữa hiệu quả của tự động hóa và sự giám sát của con người, trong đó các mô hình AI hỗ trợ việc phát hiện hoạt động đáng ngờ nhưng các quyết định cuối cùng phải có thể diễn giải và hợp lý. Những hệ lụy tiềm tàng từ việc phân loại sai trong các môi trường có rủi ro cao này nhấn mạnh sự cần thiết của các khuôn khổ đạo đức vững chắc, sự giám sát từ cơ quan quản lý, cũng như các cơ chế cho phép cá nhân bị ảnh hưởng có thể khiếu nại các quyết định do AI đưa ra.

2. Sự công bằng và thiên lệch trong các hệ thống AI

Một vấn đề nền tảng khác trong sự hợp tác giữa con người và AI là tính công bằng và sự thiên lệch trong các hệ thống AI. Các mô hình AI kế thừa thiên lệch từ dữ liệu huấn luyện, vốn có thể bắt nguồn từ bất bình đẳng lịch sử, tập dữ liệu không cân bằng hoặc các giả định thuật toán sai lầm. Hơn nữa, thiên lệch của con người cũng có thể ảnh hưởng đến dự đoán của AI khi các chuyên gia gán nhãn dữ liệu huấn luyện hoặc đưa ra quyết định theo thời gian thực cùng với AI.

Nếu không có các chiến lược giảm thiểu cẩn thận, các hệ thống AI thiên lệch có thể làm trầm trọng thêm bất bình đẳng xã hội, dẫn đến sự đối xử không công bằng trong các lĩnh vực như tuyển dụng, chẩn đoán y khoa và phát hiện gian lận. Bộ dữ liệu FiFAR giải quyết trực tiếp vấn đề này bằng cách cho phép các nhà nghiên cứu phân tích tính bình đẳng dự đoán, đảm bảo rằng các mô hình AI không gây bất lợi có hệ thống cho một số nhóm nhất định, chẳng hạn như người cao tuổi trong phát hiện gian lận tài chính.

Bằng cách cung cấp một bộ tiêu chuẩn đánh giá có cấu trúc, FiFAR giúp các nhà nghiên cứu kiểm tra liệu các mô hình AI có duy trì sự công bằng giữa các nhóm dân số đa dạng hay không, đồng thời khám phá các điều chỉnh thuật toán nhằm giảm thiểu thiên lệch mà vẫn bảo toàn hiệu suất.

3. Khả năng diễn giải và tính minh bạch của mô hình

Ngoài vấn đề công bằng, diễn giải mô hình và tính minh bạch đóng vai trò quan trọng trong việc xây dựng niềm tin đối với các quyết định do AI hỗ trợ. Nhiều mô hình AI, đặc biệt là các hệ thống dựa trên học sâu, hoạt động như một "hộp đen," khiến người dùng khó hiểu được cách một quyết định cụ thể được đưa ra. Trong các ứng dụng nhạy cảm như đánh giá rủi ro tài chính hoặc chẩn đoán y khoa, sự thiếu minh bạch có thể cản trở trách nhiệm giải trình và làm suy giảm lòng tin của người dùng.

Để giải quyết vấn đề này, nhiều kỹ thuật đã được phát

triển nhằm cải thiện khả năng diễn giải mô hình, bao gồm LIME (Giải thích mô hình cục bộ không phụ thuộc) và SHAP (Giải thích cộng tính Shapley). Những phương pháp này giúp làm rõ cách các đặc trưng đầu vào ảnh hưởng đến dự đoán của mô hình. Đặc biệt, chúng rất hữu ích trong các hệ thống Learning to Defer (L2D), nơi quyết định được phân bổ linh hoạt giữa mô hình AI và chuyên gia con người.

Bằng cách tận dụng các kỹ thuật diễn giải mô hình, các nhà nghiên cứu có thể đảm bảo rằng hệ thống L2D không chỉ tối ưu hóa hiệu suất mà còn cung cấp các giải thích rõ ràng, giúp tăng cường sự hiểu biết của con người và củng cố niềm tin vào quy trình làm việc do AI hỗ trợ.

4. Những thách thức triển khai mô hình trong thực tế

Bất chấp những tiến bộ trong công bằng AI và khả năng diễn giải mô hình, việc triển khai mô hình Learning to Defer (L2D) trong môi trường thực tế vẫn đặt ra nhiều thách thức. Khác với các điều kiện nghiên cứu được kiểm soát, triển khai thực tế đòi hỏi phải quản lý cẩn thận các giới hạn về năng lực con người, điều kiện ra quyết định luôn thay đổi và những hạn chế trong khả năng thu thập dữ liệu.

Các chuyên gia con người chỉ có thể xử lý một số lượng trường hợp hữu hạn trong một khoảng thời gian nhất định, điều này đòi hỏi các chiến lược phân phối khối lượng công việc một cách thông minh. Ngoài ra, môi trường ra quyết định trong thực tế rất động, khiến các mô hình AI phải liên tục thích ứng với các xu hướng mới, phương thức gian lận phức tạp hơn và các yêu cầu pháp lý thay đổi theo thời gian.

Một thách thức lớn khác là thu thập dữ liệu, vì việc có được dữ liệu huấn luyện chất lượng cao do con người gán nhãn thường tốn kém và mất nhiều thời gian. Bộ dữ liệu FiFAR đóng vai trò quan trọng trong việc giải quyết những thách thức này bằng cách mô phỏng các ràng buộc thực tế, giúp các nhà nghiên cứu kiểm tra và tinh chỉnh mô hình L2D trong điều kiện phản ánh sự phức tạp ngoài đời thực.

Nhờ đó, FiFAR giúp phát triển các hệ thống hợp tác AI - con người có tính thích ứng cao, hiệu quả, và công bằng hơn, nhằm triển khai trong các ứng dụng quan trọng với độ tin cậy cao hơn.

IV. Phân tích và thực nghiệm

1. Điểm mạnh của FIFAR

Bộ dữ liệu FiFAR được thiết kế để giải quyết những hạn chế hiện có trong nghiên cứu về Learning to Defer (L2D), đặc biệt trong lĩnh vực phát hiện gian lận tài chính. Một trong những điểm mạnh cốt lõi của FiFAR là khả năng mô phỏng môi trường thực tế bằng cách cung cấp các dự đoán từ 50 chuyên gia giả lập có mức độ phức tạp cao, giúp tái tạo sự đa dạng trong hành vi con người. Ngoài ra, FiFAR tích hợp các ràng buộc về năng lực xử lý của con người, phản ánh thực tế rằng chuyên gia chỉ có thể xử lý một số lượng trường hợp giới hạn trong một khoảng thời gian nhất định. Điều này cho phép thử nghiệm các thuật toán phân bổ quyết định trong điều kiện gần với thực tế, thúc đẩy sự phát triển của các hệ thống AI - con người có tính ứng dụng cao hơn.

Một trong những thách thức lớn trong nghiên cứu L2D là thiếu dữ liệu thực tế về quyết định của con người, do chi phí thu thập và gán nhãn dữ liệu rất cao. FiFAR khắc phục vấn đề này bằng cách tạo ra một tập dữ liệu tổng hợp với các chuyên gia giả lập có độ chính xác, mức độ phụ thuộc vào đặc trưng và thiên lệch có kiểm soát. Điều này giúp các nhà nghiên cứu kiểm tra tính công bằng của mô hình AI, đồng thời tối ưu hóa hiệu suất thông qua các thuật toán học tăng cường. Bên cạnh đó, FiFAR còn hỗ trợ đánh giá khả năng thích ứng của mô hình trước các điều kiện thay đổi, như sự biến động trong khả năng làm việc của con người hoặc sự thay đổi trong quy định pháp lý.

FiFAR không chỉ giúp kiểm định các thuật toán hiện có mà còn đóng vai trò là một tiêu chuẩn đánh giá (benchmark) quan trọng trong nghiên cứu về hợp tác AI - con người. Bộ dữ liệu này cho phép so sánh hiệu suất giữa các mô hình khác nhau trong điều kiện ràng buộc thực tế, giúp phát triển các hệ thống AI công bằng hơn và có khả năng hỗ trợ ra quyết định tốt hơn. Nhờ đó, FiFAR góp phần thúc đẩy nghiên cứu về sự cân bằng giữa công bằng và hiệu suất (fairness-performance balance), một yếu tố then chốt trong việc triển khai AI vào các ứng dụng quan trọng như phát hiện gian lận tài chính.

2. Hạn chế của nghiên cứu

Mặc dù FiFAR mang lại nhiều lợi ích trong nghiên cứu về Learning to Defer (L2D), nhưng nó vẫn tồn tại một số hạn chế quan trọng. Một trong những thách thức lớn nhất là việc dựa vào chuyên gia tổng hợp (synthetic experts) thay vì dữ liệu thực tế. Các mô hình chuyên gia tổng hợp trong FiFAR được thiết kế để mô phỏng hành vi con người, nhưng không thể hoàn toàn thay thế dữ liệu từ các chuyên gia thực tế. Điều này có thể dẫn đến sai lệch trong việc đánh giá hiệu suất của hệ thống AI - con người, do hành vi giả lập có thể không phản ánh đầy đủ sự phức tạp và không nhất quán của con người trong thực tế.

Một hạn chế quan trọng khác là thiếu dữ liệu thực tế về quyết định của con người, đặc biệt là trong các tình huống phức tạp như phát hiện gian lận tài chính. Việc thu thập dữ liệu thực tế là rất tốn kém và mất thời gian, khiến nhiều nghiên cứu phải dựa vào dữ liệu tổng hợp. Tuy nhiên, dữ liệu tổng hợp có thể không phản ánh chính xác các yếu tố như tâm lý, kinh nghiệm, hoặc sai sót của con người, điều này ảnh hưởng đến tính khả thi khi triển khai mô hình vào thực tế.

Một vấn đề khác khi sử dụng dữ liệu tổng hợp là khả năng khuếch đại thiên kiến trong mô hình AI. Trong FiFAR, các chuyên gia giả lập được thiết kế với mức độ phụ thuộc vào đặc trưng và mô hình AI có kiểm soát, nhưng điều này có thể vô tình củng cố các thiên lệch có sẵn trong dữ liệu huấn luyện. Nếu một chuyên gia giả lập có thiên lệch chống lại một nhóm nhất định (ví dụ: người cao tuổi trong các quyết định về gian lận tài chính), mô hình AI có thể học và tiếp tục duy trì hoặc thậm chí khuếch đại thiên kiến này. Điều này đặt ra thách thức trong việc đảm

Bång 1: Baseline Results. Intervals denote standard deviation.

Scenario Properties					ModelOnly		ReL		ReLgreedy		ReLlinear	
Pool	Batch size	Deferral rate	Absence rate	σ_d	Loss	PE	Loss	PE	Loss	PE	Loss	PE
all	250	0.2	0.0	0.2	918	0.33	753±12	0.29	780±10	0.31	780±9	0.32
all	250	0.2	0.0	0.0	918	0.33	755±14	0.30	781±8	0.31	789±8	0.32
all	250	0.2	0.5	0.2	918	0.33	760±11	0.29	788±9	0.31	782±9	0.32
all	250	0.2	0.5	0.0	918	0.33	768±10	0.29	788±11	0.31	786±10	0.32
all	250	0.5	0.0	0.2	918	0.33	746±17	0.29	788±7	0.34	766±9	0.36
all	250	0.5	0.0	0.0	918	0.33	759±14	0.29	790±4	0.34	765±13	0.36
all	250	0.5	0.5	0.2	918	0.33	756±13	0.29	779±7	0.32	782±8	0.36
all	250	0.5	0.5	0.0	918	0.33	754±11	0.29	783±6	0.32	783±5	0.36
all	5000	0.2	0.0	0.2	918	0.33	752±8	0.30	780±4	0.32	779±5	0.33
all	5000	0.2	0.0	0.0	918	0.33	752±12	0.30	778±3	0.32	782±5	0.33
all	5000	0.2	0.5	0.2	918	0.33	762±10	0.30	778±12	0.31	773±4	0.33
all	5000	0.2	0.5	0.0	918	0.33	753±9	0.30	776±11	0.31	776±3	0.33
all	5000	0.5	0.0	0.2	918	0.33	749±8	0.29	774±6	0.34	768±6	0.36
all	5000	0.5	0.0	0.0	918	0.33	750±11	0.29	776±8	0.34	768±1	0.36
all	5000	0.5	0.5	0.2	918	0.33	759±12	0.29	774±7	0.32	780±8	0.37
all	5000	0.5	0.5	0.0	918	0.33	758±8	0.29	773±6	0.33	781±7	0.37

Bång 2: Varying Expert Pool Results.

Scenario Properties				ReL		ReLgreedy		ReLlinear	
Pool	Batch Size	Deferral Rate	Absence Rate	Loss	PE	Loss	PE	Loss	PE
agreeing	250	0.2	0.0	813±8	0.37	873±7	0.37	810±3	0.34
agreeing	5000	0.2	0.0	816±7	0.37	875±4	0.37	808±5	0.34
agreeing	5000	0.5	0.0	814±12	0.39	905±3	0.40	784±3	0.36
sparse	250	0.2	0.0	766±9	0.29	770±6	0.31	755±6	0.31
sparse	250	0.5	0.0	752±11	0.28	738±8	0.31	737±11	0.34
sparse	5000	0.2	0.0	752±4	0.29	767±5	0.31	738±3	0.32
sparse	5000	0.5	0.0	764±11	0.29	758±4	0.32	737±5	0.34
standard	250	0.2	0.0	742±13	0.30	782±12	0.32	788±7	0.33
standard	250	0.5	0.0	739±9	0.31	773±9	0.33	782±6	0.34
standard	5000	0.2	0.0	739±6	0.31	773±1	0.32	773±4	0.33
standard	5000	0.5	0.0	731±12	0.31	757±2	0.33	777±4	0.35
unfair	250	0.2	0.0	736±8	0.22	721±6	0.24	714±1	0.25
unfair	250	0.5	0.0	722±9	0.19	708±3	0.21	687±8	0.23
unfair	5000	0.2	0.0	724±11	0.23	726±2	0.25	711±2	0.26
unfair	5000	0.5	0.0	712±7	0.20	712±3	0.22	682±5	0.24

bảo tính công bằng khi áp dụng mô hình vào thực tế.

Để khắc phục những hạn chế này, cần có các phương pháp đánh giá và hiệu chỉnh thiên lệch trong dữ liệu tổng hợp, cũng như tích hợp dữ liệu thực tế khi có thể. Ngoài ra, nghiên cứu trong tương lai có thể tập trung vào việc phát triển các kỹ thuật tạo dữ liệu tổng hợp có kiểm soát tốt hơn, nhằm đảm bảo rằng mô hình AI không chỉ đạt hiệu suất cao mà còn duy trì tính công bằng trong quyết định.

3. Thiết lập thí nghiệm

Tiêu chí Neyman-Pearson và Hàm mất mát nhạy cảm với chi phí Như đã đề cập trong Mục 3.1, mục tiêu tối ưu hóa là tối đa hóa recall tại mức **FPR = 5%** (Tiêu chí Neyman-Pearson). Khi đánh giá một tập hợp các phân bổ, giá trị của FPR có thể không giống nhau giữa các thí nghiệm. Điều này gây khó khăn trong việc so sánh trực tiếp recall của các thuật toán khác nhau (tức là, nếu hai phương pháp đạt được cùng một recall, phương pháp có FPR thấp hơn được ưu tiên hơn). Mục tiêu tối ưu hóa này ngầm thể hiện một sự đánh đổi giữa chi phí phân loại sai của lỗi FP và FN, tức là một bài toán nhạy cảm với chi phí.

Để đánh giá hiệu suất, ta có thể sử dụng một hàm mất mát nhạy cảm với chi phí như sau:

$$L = \lambda N(FP) + N(FN), \tag{3}$$

trong đó λ được định nghĩa là:

$$\lambda = \frac{t}{1-t}, \tag{4}$$

trong đó $N(FP)$ và $N(FN)$ lần lượt là số lượng lỗi FP và FN.

Tham số λ xác định mối quan hệ giữa chi phí của một lỗi FP và một lỗi FN. Chúng ta cần xác định mối quan hệ giữa tiêu chí Neyman-Pearson và giá trị của λ. Theo Elkan [2], giá trị ngưỡng tối ưu t của một bộ phân loại nhị phân và chi phí phân loại sai có mối quan hệ theo phương trình trên. Khi huấn luyện bộ phân loại học máy, ngưỡng của nó được chọn phù hợp với tiêu chí Neyman-Pearson, do đó ta đặt λ dựa trên ngưỡng t của mô hình.

*4. **Baseline***

Khi tìm kiếm các base line L2D khả thi, chúng tôi không tìm thấy phương pháp hiện tại nào có thể tính đến các ràng buộc dung lượng cá nhân. Do đó, chúng tôi cung cấp ba đường cơ sở như sau.

Học từ chối (Rejection Learning - ReL).Trong triển khai này, chúng tôi sử dụng điểm số của mô hình như một thước đo độ tin cậy của mô hình. Để áp dụng học từ chối theo lô, trong phạm vi các ràng buộc dung lượng, trước tiên chúng tôi sắp xếp các trường hợp trong lô theo thứ tự giảm dần của điểm số mô hình. 5% trường hợp có điểm số cao nhất sẽ tự động bị từ chối (dự đoán dương tính). Sau đó, các trường hợp tiếp theo được gán ngẫu nhiên cho các chuyên gia trong nhóm của chúng tôi cho đến khi đáp ứng giới hạn dung lượng của họ. Tất cả các trường hợp còn lại, với điểm số mô hình thấp nhất, sẽ được phân loại là chấp nhận (dự đoán âm tính).

Học từ chối có nhận thức về chuyên gia (Human Expertise Aware Rejection Learning) Trong phiên bản học từ chối này, thay vì gán ngẫu nhiên các trường hợp bị từ chối cho nhóm chuyên gia, chúng tôi cố gắng mô hình hóa hành vi của từng cá nhân để tối ưu hóa việc phân bổ. Để làm điều này, chúng tôi huấn luyện một mô hình LightGBM dựa trên các đặc trưng của trường hợp và *expert_id* để dự đoán xem chuyên gia có đưa ra dự đoán sai dương tính (FP), sai âm tính (FN), đúng dương tính (TP) hay đúng âm tính (TN) hay không. Đối với mỗi trường hợp, chúng tôi có thể dự đoán xác suất chuyên gia sẽ mắc lỗi FP hoặc FN, được ký hiệu lần lượt là $\hat{P}(FP)$ và $\hat{P}(FN)$.

Sau đó, chúng tôi tính toán mất mát dự đoán liên quan đến việc chuyển trường hợp X_i cho chuyên gia e:

$$L(X_i, e) = \lambda \hat{P}(FP) + \hat{P}(FN), \tag{5}$$

Chúng tôi trình bày hai phiên bản của thuật toán này:

- **Tham lam (ReLgreedy)**: Thuật toán xử lý từng trường hợp trong lô, gán từng trường hợp cho chuyên gia có mất mát dự đoán thấp nhất. Nếu một lần gán vi phạm ràng buộc dung lượng, thuật toán sẽ thử gán cho chuyên gia có mất mát thấp thứ hai, và tiếp tục như vậy cho đến khi tất cả các ràng buộc được đáp ứng.
- **Lập trình tuyến tính nguyên (ReLlinear)**: Trong phương pháp này, chúng tôi tối thiểu hóa tổng mất mát trên toàn bộ lô bằng cách giải một bài toán lập trình tuyến tính, có các ràng buộc dung lượng, nhằm tìm ra phân bổ tối ưu trên toàn bộ lô.

Trong bối cảnh công bằng, chúng tôi muốn đảm bảo rằng xác suất từ chối sai một đơn đăng ký hợp lệ là độc lập với giá trị thuộc tính nhạy cảm. Do đó, chúng tôi đo tỷ lệ giữa các FPR trong từng nhóm tuổi, hay còn gọi là **công bằng dự đoán** (Predictive Equality - PE) [**?**]. Tỷ lệ này được tính bằng cách chia FPR của nhóm có FPR quan sát thấp nhất cho FPR của nhóm có FPR cao nhất.

5. Kết quả

Trong Bảng 1, chúng tôi trình bày kết quả cho các mô hình L2D cơ sở được thảo luận, cũng như hệ thống "Chỉ có mô hình". Hàm mất mát được tính theo Phương trình 1, với N(FP) và N(FN) được đếm trên tập kiểm tra. Chúng tôi có thể thấy cách kết quả của từng mô hình L2D cơ sở thay đổi theo môi trường hợp tác giữa con người và AI (Thuộc tính Kịch bản) trên các hàng của bảng, đối với chỉ số hiệu suất (mất mát nhạy cảm với chi phí) và chỉ số công bằng (Bình đẳng dự đoán).

Chúng tôi quan sát thấy rằng trong tất cả các kịch bản được xem xét, học từ chối (rejection learning) hoạt động tốt nhất, bất chấp nỗ lực của chúng tôi trong việc mô hình hóa hành vi con người. Điều này có thể là do số lượng FN và TP trong dữ liệu huấn luyện thấp, dẫn đến ước lượng xác suất kém và xếp hạng sai xác suất mắc lỗi của chuyên gia đối với một mẫu cụ thể.

Trong Phụ lục, Mục D, Bảng 6 cho thấy rằng các phương

pháp của chúng tôi chủ yếu giảm thiểu lỗi FP, dẫn đến tỷ lệ FP (FPR) thấp hơn nhưng cũng làm giảm recall. Việc giảm thiểu False Positives cũng giúp tăng bình đẳng dự đoán, cho thấy rằng mô hình chuyên gia của chúng tôi đã học được rằng các chuyên gia có xu hướng mắc lỗi FP nhiều hơn trên các đơn đăng ký của khách hàng lớn tuổi.

Một sự thay đổi đáng kể có thể thấy trong kết quả của phương pháp ReLlinear. Trong khi sự vắng mặt của chuyên gia dường như không ảnh hưởng đáng kể đến mất mát trong các kịch bản có tỷ lệ hoãn quyết định (deferral rate) là 20%, thì trong các trường hợp có tỷ lệ hoãn quyết định 50%, việc thiếu vắng chuyên gia dường như làm tăng đáng kể mất mát. Điều này minh họa tầm quan trọng của việc kiểm tra hệ thống dưới nhiều điều kiện khác nhau.

Trong Bảng 2, chúng tôi giới thiệu sự thay đổi trong nhóm chuyên gia sẵn có. Tại đây, chúng tôi có thể thấy rằng ReLlinear vượt trội hơn ReL khi nhóm chuyên gia chỉ bao gồm các chuyên gia đồng thuận, chuyên gia rời rạc hoặc chuyên gia không công bằng. Điều này có thể là do những chuyên gia này dễ mô hình hóa hơn, vì họ có một đặc trưng chi phối rõ ràng hoặc có sự phụ thuộc vào đặc trưng đơn giản hơn. Điều này minh họa tầm quan trọng của việc xem xét mức độ phức tạp khác nhau của hành vi con người khi đánh giá các hệ thống hợp tác giữa con người và AI (HAIC).

V. Conclusion

Trong bài báo này, chúng tôi giới thiệu FiFAR Dataset, một nguồn dữ liệu mới được thiết kế để hỗ trợ đánh giá các thuật toán Learning to Defer (L2D) trong điều kiện thực tế. Để chứng minh tính ứng dụng, chúng tôi đã kiểm tra ba phương pháp L2D cơ bản trên 300 kịch bản khác nhau, giúp đánh giá toàn diện hiệu suất của chúng trong thực tế.

Mặc dù có những đóng góp quan trọng, nghiên cứu của chúng tôi có một hạn chế đáng kể: các mô hình baseline được thử nghiệm chưa tích hợp các phương pháp L2D đã có trong tài liệu trước đây, do phần lớn các phương pháp này không xét đến ràng buộc dung lượng.

Chúng tôi nhấn mạnh rằng, chuyên gia tổng hợp là một công cụ hữu ích cho nghiên cứu nhưng không thể thay thế chuyên gia thực trong các hệ thống Human-AI Collaboration (HAIC). Việc phát triển các mô hình AI có thể triển khai trong thực tế đòi hỏi dữ liệu từ chuyên gia thật, vì đây là yếu tố quan trọng để đảm bảo độ tin cậy của các quyết định hỗ trợ bởi AI. Một mối quan ngại là việc sử dụng chuyên gia tổng hợp có thể làm giảm nhu cầu gán nhãn dữ liệu thủ công, đặc biệt trên các nền tảng lớn như MTurk. Tuy nhiên, trong các lĩnh vực chuyên môn cao, dữ liệu từ các dịch vụ gán nhãn cộng đồng có thể không đảm bảo chất lượng, khiến việc sử dụng chuyên gia tổng hợp trở thành một lựa chọn cần thiết. Nếu có dữ liệu dự đoán từ chuyên gia thực, chúng nên được ưu tiên hơn dữ liệu tổng hợp, vì chúng phản ánh chính xác hơn mô hình ra quyết định của con người.

Ngoài ra, chúng tôi cũng đề cập đến một thách thức quan trọng khi sử dụng chuyên gia tổng hợp: sự khuếch đại thiên vị. Do chuyên gia tổng hợp hoạt động dựa trên mối quan hệ được xác định trước giữa các đặc trưng và xác suất lỗi, chúng có nguy cơ củng cố các thiên vị đối với các nhóm đối tượng được bảo vệ. Cụ thể, nếu một số đặc trưng nhất định ảnh hưởng mạnh đến tỷ lệ dự đoán dương tính sai (false positives), mô hình có thể vô tình gây ra sự đối xử không công bằng với một số nhóm nhân khẩu học. Để giảm thiểu vấn đề này, cần đánh giá kỹ lưỡng tính công bằng dự đoán (Predictive Equality) khi sử dụng chuyên gia tổng hợp nhằm đảm bảo công bằng trong ra quyết định.

Cuối cùng, nghiên cứu của chúng tôi nhấn mạnh tầm quan trọng của việc phát triển các mô hình chuyên gia tổng hợp tinh vi hơn, giúp thu hẹp khoảng cách giữa nghiên cứu AI và ứng dụng thực tế. Bằng cách giải quyết các ràng buộc dung lượng và các vấn đề đạo đức, chúng tôi hy vọng sẽ thúc đẩy lĩnh vực Human-AI Collaboration, hướng đến việc xây dựng các hệ thống ra quyết định bằng AI công bằng, minh bạch và hiệu quả.

Tài liệu

[1] Chen, C., & Yao, M. Z. (2022). Strategic use of immersive media and narrative message in virtual marketing: Understanding the roles of telepresence and transportation. *Psychology and Marketing, 39*(3), 524–542. https://doi.org/10.1002/mar.21630

[2] Charles Elkan. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI 2001)*, pages 973–978. Morgan Kaufmann, 2001.