RESEARCH ARTICLE

PHÂN LỚP DỮ LIỆU MẤT CÂN BẰNG TRONG BÀI TOÁN DỰ ĐOÁN THUÊ BAO RỜI BỎ NHÀ MẠNG DỰA VÀO GIẢI THUẬT RỪNG NGẪU NHIÊN CẢI TIẾN

Đinh Minh Hòa, Dương Tuấn Anh

Khoa Công Nghệ Thông Tin, Trường Đại Học Ngoại Ngữ -Tin Học, Thành Phố Hồ Chí Minh

hoa.net@gmail.com, anhdt@huflit.edu.vn

TÓM TẮT— Trong lãnh vực viễn thông, việc thuê bao rời bỏ nhà mạng là sự cố rất đáng quan tâm vì vấn đề này có thể ảnh hưởng đến lợi nhuận của công ty. Tuy nhiên, đặc điểm dữ liệu mất cân bằng trong bài toán dự đoán thuê bao rời bỏ nhà mạng gây khó khăn cho việc phát triển một mô hình phân lớp hiệu quả để giải quyết bài toán này. Trong nghiên cứu này, chúng tôi thử áp dụng giải thuật Rừng ngẫu nhiên có điều chỉnh hàm chi phí (cost-sensitive weighted random forest - CSWRF), vốn đã thành công trong bài toán phát hiện gian lận thẻ tín dụng (credit card fraud detection) để giải quyết vấn đề dữ liệu mất cân bằng trong bài toán dự đoán thuê bao rời bỏ nhà mạng. Ngoài ra, chúng tôi so sánh hiệu quả của giải thuật CSWRF với cách tiếp cận lấy mẫu dữ liệu: kết hợp giải thuật Rừng ngẫu nhiên với kỹ thuật lấy mẫu tăng SMOTE (Synthetic Minority Oversampling Technique). Kết quả thực nghiệm trên hai bộ dữ liệu mẫu cho thấy đối với bài toán dự đoán thuê bao rời bỏ nhà mạng, vốn là bài toán mất cân bằng dữ liệu, hiệu quả phân lớp của giải thuật CSWRF thuộc cách tiếp cận điều chỉnh hàm chi phí (cost-sensitive learning) tốt hơn phương pháp SMOTE kết hợp giải thuật Rừng ngẫu nhiên.

Từ khóa— dữ liệu mất cân bằng, dự đoán thuê bao rời bỏ nhà mạng, rừng ngẫu nhiên, cách tiếp cận điều chỉnh hàm chi phí, lấy mẫu tăng SMOTE.

I. GIỚI THIỆU

Quản lý mối quan hệ khách hàng (customer relationship management - CRM) là cách tiếp cận chiến lược nhằm phát triển những mối quan hệ dài lâu, hữu ích với những khách hàng trung thành của công ty. Do tình trạng thị trường bão hòa và cạnh tranh mạnh mẽ, ngày càng có nhiều công ty nhận ra tầm quan trọng của việc quản lý mối quan hệ khách hàng và nguồn dữ liệu của các hệ thống CRM rất hữu ích để ra quyết định về mặt quản lý. Ngày nay, sự phát triển nhanh chóng của các kỹ thuật học máy đem lại những cơ hội để hiểu biết sâu hơn về khách hàng và xây dựng những hệ thống quản lý mối quan hệ khách hàng hiệu quả.

Dự đoán thuê bao rời bỏ nhà mạng (telecom customer churn prediction) như là một phần của công tác quản lý mối quan hệ khách hàng đã trở thành một chủ đề đáng quan tâm hiện nay. Trong lĩnh vực viễn thông điện thoại di động, thuật ngữ "rời bỏ" nói về sự ra đi của những thuê bao dịch chuyển từ nhà cung cấp này sang nhà cung cấp khác trong một khoảng thời gian nhất định. Do việc giữ lại được những khách hàng hiện hành thì có lợi hơn việc thu hút thêm những khách hàng mới, nên công việc xây dựng một mô hình dự đoán thuê bao rời bỏ nhà mạng hiệu quả là rất quan trọng để sớm nhận dạng những khách hàng có khuynh hướng rời bỏ công ty.

Các công trình nghiên cứu liên quan đến dự đoán thuê bao rời bỏ nhà mạng thường sử dụng các kỹ thuật khai phá dữ liệu, như mạng nơ ron, gom cụm, cây quyết định, giải thuật k-lân cận gần nhất, hồi quy logistic, máy vector hỗ trợ, và tổ hợp (ensemble) nhiều phương pháp học máy để đem lại sự dự báo chính xác [1].

Tuy nhiên, trong bài toán dự đoán thuê bao rời bỏ nhà mạng, tập dữ liệu được dùng để huấn luyện mô hình phân lớp thường *mất cân bằng* (imbalanced), tức là nhóm thuê bao không rời bỏ chiếm đại đa số trong tập dữ liệu. Các phương pháp phân lớp truyền thống thường không thể ứng phó với vấn đề này vì chúng có khuynh hướng phân lớp mẫu thử về *lớp đa số* (majority class) và xem các mẫu thuộc *lớp thiểu số* (minority class) như là nhiễu hay là những điểm ngoại biên (outlier). Khuynh hướng này đã khiến cho công tác phân lớp trong bài toán dự đoán thuê bao rời bỏ nhà mạng trở nên không hiệu quả [2]. Do đó xử lý hiệu quả vấn đề dữ liệu mất cân bằng trong bài toán phân lớp rất quan trọng sẽ giúp cải tiến độ chính xác phân lớp trong nhiều ứng dụng có ý nghĩa thực tiễn như chẩn đoán bệnh hiếm có, phát hiện gian lận, phát hiện xâm nhập mạng trái phép, dự đoán thuê bao rời bỏ công ty, dự đoán công ty phá sản, dự đoán nhân viên bỏ việc.

Trong những năm gần đây, nhiều giải pháp đã được đề xuất để giải quyết vấn đề phân lớp với dữ liệu mất cân bằng bao gồm các giải thuật phân lớp và các phương pháp tổ hợp (ensemble method). Các phương pháp này được chia làm hai loại: nhóm phương pháp *lấy mẫu dữ liệu* (data sampling) nhằm cân bằng dữ liệu trước giai đoạn huấn luyện và các phương pháp *điều chỉnh hàm chi phí* (cost-sensitive learning) nhằm chỉnh sửa cách tính chi phí tương đối của những lỗi sai phân lớp trong giai đoạn huấn luyện. Gần đây, một công trình khảo sát bằng thực nghiệm [3] chỉ ra rằng cách tiếp cận điều chỉnh hàm chi phí đem lại một hiệu quả cao hơn cách tiếp cận lấy mẫu dữ liệu khi xử lý vấn đề dữ liệu mất cân bằng trong bài toán dự đoán thuê bao rời bỏ nhà mạng.

Được gợi cảm hứng từ công trình [4] của Devi và các cộng sự (2019) đề xuất giải thuật Rừng ngẫu nhiên có điều chỉnh hàm chi phí (Cost-Sensitive Weighted Random Forest – CSWRF) để giải quyết bài toán *phát hiện gian lận thẻ tín dụng* (credit card fraud detection). Trong nghiên cứu này, chúng tôi thử ứng dụng giải thuật CSWRF cho một bối cảnh khác: bài toán dự đoán thuê bao rời bỏ nhà mạng, mà có cùng đặc điểm là dữ liệu mất cân bằng. Ngoài ra, chúng tôi tiến hành so sánh hiệu quả phân lớp của giải thuật CSWRF với phương pháp kết hợp kỹ thuật lấy mẫu tăng SMOTE và giải thuật Rừng ngẫu nhiên thuần túy cho bài toán dự đoán thuê bao rời bỏ nhà mạng . Kết quả thực nghiệm trên hai bộ dữ liệu mẫu chuẩn Telco và Cell2Cell cho thấy với bài toán dự đoán thuê bao rời

bỏ nhà mạng, hiệu quả phân lớp của giải thuật CSWRF tốt hơn hiệu quả phân lớp của phương pháp SMOTE kết hợp với giải thuật Rừng ngẫu nhiên.

Phần tiếp theo của bài báo được tổ chức như sau: Mục II giới thiệu các công trình liên quan về xử lý dữ liệu mất cân bằng trong bài toán dự đoán thuê bao rời bỏ nhà mạng; Mục III mô tả cách tiếp cận đề xuất là sử dụng giải thuật Rừng ngẫu nhiên có điều chỉnh hàm chi phí (CSWRF); Mục IV trình bày kết quả thực nghiệm so sánh hiệu quả phân lớp của giải thuật CSWRF với phương pháp kết hợp kỹ thuật lấy mẫu tăng SMOTE và giải thuật Rừng ngẫu nhiên; Mục V nêu một vài kết luận và các hướng phát triển của đề tài.

II. CÁC CÔNG TRÌNH LIÊN QUAN

A. PHƯƠNG PHÁP HỌC MÁY ĐỂ DỰ ĐOÁN THUÊ BAO RỜI BỎ NHÀ MẠNG

Để hiểu biết các nghiên cứu xây dựng những mô hình khác nhau cho bài toán dự đoán thuê bao rời bỏ nhà mạng, tiểu mục này điểm qua một số công trình nghiên cứu liên quan như sau:
Sharma và Panigrahi (2013)đề xuất một mô hình mạng nơ ron cho bài toán dự đoán thuê bao rời bỏ nhà mạng [5]. Zhao và các cộng sự, năm 2005 [6] đề xuất máy vector hỗ trợ một lớp (one-class support vector machines) để dự đoán thuê bao rời bỏ nhà mạng. Chất lượng dự đoán của mô hình máy vector hỗ trợ được so sánh với mạng nơ ron nhân tạo, cây quyết định và phương pháp Naive Bayes. Zhang và các cộng sự, năm 2007 [7] sử dụng mô hình lai giữa giải thuật k-lân cận gần nhất và hồi quy logistic để dự đoán thuê bao rời bỏ nhà mạng. Một nghiên cứu khác (Lu và các cộng sự, 2014 [8]) sử dụng giải thuật boosting để xây dựng mô hình dự đoán thuê bao rời bỏ nhà mạng. Estes và Mendes-Moreia (2016) [9] so sánh hiệu năng của sáu giải thuật phân lớp, gồm k-lân cận gần nhất, Naïve-Bayes, Cây Quyết Định C4.5, AdaBoost, mạng nơ ron và giải thuật Rừng ngẫu nhiên cho bài toán dự đoán thuê bao rời bỏ nhà mạng.
Các công trình nghiên cứu nêu trên, phần lớn chỉ tập trung sử dụng một phương pháp khai phá dữ liệu duy nhất như phân lớp hay gom cụm để phân tích dữ liệu dự đoán thuê bao rời bỏ nhà mạng. Chỉ có một vài công trình áp dụng nhiều hơn một phương pháp dựa vào phân lớp hay gom cụm, thí dụ, Li và Deng (2012) đã kết hợp gom cụm và cây quyết định để dự đoán thuê bao rời bỏ nhà mạng cho công ty China Telecom [10].
Tuy nhiên, tất cả các công trình nêu trên có một nhược điểm chung là chưa để ý đến việc giải quyết vấn đề mất cân bằng dữ liệu khi phân lớp, do đó hiệu quả phân lớp của chúng còn thấp, chưa đáp ứng được yêu cầu thực tế của bài toán dự đoán thuê bao rời bỏ nhà mạng.

B. XỬ LÝ MẤT CÂN BẰNG DỮ LIỆU

Trong những năm gần đây, nhiều cách tiếp cận đã được đề xuất để giải quyết vấn đề mất cân bằng dữ liệu khi phân lớp và từ đó đã nâng cao được hiệu quả phân lớp cho nhiều ứng dụng gặp phải dữ liệu mất cân bằng. Các phương pháp này được gom thành hai nhóm như sau:

1. LẤY MẪU DỮ LIỆU

Các phương pháp *lấy mẫu dữ liệu* (data sampling) lại được chia làm ba loại: *lấy mẫu giảm* (under-sampling) và *lấy mẫu tăng* (over-sampling) và sự kết hợp giữa lấy mẫu giảm với lấy mẫu tăng. Lấy mẫu giảm loại bỏ bớt một số mẫu thuộc lớp đa số trong khi lấy mẫu tăng tạo ra thêm một số mẫu mới cho lớp thiểu số. Cả hai phương pháp lấy mẫu dữ liệu đều nhắm đến việc giảm bớt hiệu ứng xấu của phân bố lớp bị lệch (skew distribution) trong quá trình huấn luyện.

Một số phương pháp lấy mẫu dữ liệu được ưa chuộng có thể kể như sau. Chawla và các cộng sự năm 2002 [11] đề xuất một phương pháp lấy mẫu tăng SMOTE (Synthetic Minority Oversampling Technique). Có hai phương pháp lấy mẫu giảm khá đơn giản nhưng rất hiệu quả là phương pháp Lấy mẫu giảm ngẫu nhiên (Random Under-Sampling - RUS) mà thực hiện việc loại bỏ một cách ngẫu nhiên một số mẫu thuộc lớp đa số (do Tahir và các cộng sự đề xuất năm 2009 [12]) và phương pháp lấy mẫu giảm dựa vào gom cụm để thu giảm một số mẫu trong lớp đa số (do Lin và các cộng sự đề xuất năm 2017 [13]). Theo bài báo tổng quan (review) [14], Haixiang và các cộng sự, năm 2017, nhận xét rằng các phương pháp lấy mẫu tăng được sử dụng phổ biến hơn các phương pháp lấy mẫu giảm và khi số lượng mẫu của lớp thiểu số khá nhỏ so với số lượng của lớp đa số thì phương pháp lấy mẫu tăng SMOTE là một lựa chọn thích hợp.

2. ĐIỀU CHỈNH HÀM CHI PHÍ

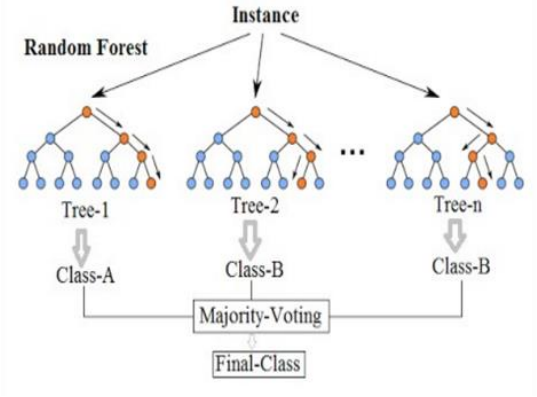
Trong nhiều ứng dụng khai phá dữ liệu, kể cả dự đoán thuê bao rời bỏ nhà mạng, những mẫu thuộc lớp thiểu số đóng vai trò rất quan trọng. Hàm chi phí (cost function) hoặc hàm mất mát (loss function) của giải thuật phân lớp mà không quan tâm đến điều này sẽ không thể hoạt động hiệu quả trong những tình huống như vậy. Một cách tiếp cận để giải quyết vấn đề là áp dụng những phương pháp *điều chỉnh hàm chi phí* (cost-sensitive learning). Những phương pháp này gán chi phí cao cho sự phân lớp sai các mẫu thuộc lớp thiểu số so với các mẫu thuộc lớp đa số và do đó cố gắng tối thiểu hóa những lỗi chi phí cao. Những phương pháp điều chỉnh hàm chi phí có thể tích hợp vào khía cạnh dữ liệu hay khía cạnh giải thuật hoặc cả hai [15].

Vài phương pháp điều chỉnh hàm chi phí tiêu biểu có thể liệt kê như sau. Lin và các cộng sự năm 2017 [16] đề xuất một phương pháp điều chỉnh hàm chi phí sử dụng Focal Loss, một độ đo lỗi dựa vào entropy, để giải quyết vấn đề mất cân bằng dữ liệu trong bài toán nhận dạng đối tượng (object detection) bằng bộ nhận dạng

RetinaNet. Wang và các cộng sự năm 2020 [17] đề xuất hai phương pháp: sử dụng Focal loss vào giải thuật XGBoost (một giải thuật tổ hợp bộ phân lớp) [18] và sử dụng Cross-entropy loss có trọng số vào giải thuật XGBoost để giải quyết vấn đề mất cân bằng dữ liệu. Công trình [17] thử nghiệm hai phương pháp đề xuất trên bộ dữ liệu chẩn đoán bệnh Parkinson, là bộ dữ liệu mất cân bằng với tỉ số 1/3, và nhận thấy cả hai phương pháp này đều đem lại hiệu quả phân lớp khá tốt và độ tốt của chúng ngang bằng nhau. Tuy nhiên cho đến nay vẫn chưa có công trình nghiên cứu nào so sánh hiệu quả của hai phương pháp XGBoost sử dụng Focal loss và XGBoost sử dụng weighted cross entropy loss với các phương pháp điều chỉnh hàm chi phí dựa vào các giải thuật phân lớp tổ hợp khác như AdaBoost hay Rừng Ngẫu Nhiên.

C. GIẢI THUẬT RỪNG NGẪU NHIÊN VÀ CÁC PHIÊN BẢN MỞ RỘNG

Giải thuật *Rừng ngẫu nhiên* (Random Forest), được phát triển bởi Breiman năm 2001 [19], là một tổ hợp (ensemble) những cây quyết định. Rừng ngẫu nhiên là một giải thuật phân lớp nếu giá trị đầu ra là giá trị rời rạc và là một giải thuật dự báo nếu giá trị đầu ra là một giá trị liên tục. Rừng ngẫu nhiên làm việc với hai giai đoạn. Đầu tiên, giải thuật tạo ra những tập huấn luyện con (training subset) từ tập huấn luyện gốc dựa vào kỹ thuật lấy mẫu boostrap và sau đó xây dựng những cây quyết định thành phần từ những tập huấn luyện con. Trong quá trình xây dựng một cây quyết định thành phần, mỗi nút trong cây được tách dựa vào một tập con các đặc trưng được chọn từ tập đặc trưng ban đầu. Trong giai đoạn hai, cho một mẫu thử, giải thuật sẽ thực hiện một chiến lược bầu phiếu đơn giản: gán mẫu thử vào nhãn lớp mà chiếm đa số phiếu bầu từ tất cả những cây quyết định thành phần. Hình 1 minh hoạ cách vận hành của Rừng ngẫu nhiên.



Hình 1. *Minh họa cách vận hành của Rừng ngẫu nhiên*

Rừng ngẫu nhiên phân lớp với hiệu quả cao hơn các giải thuật phân lớp khác như k-lân cận gần nhất, Naïve-Bayes, cây quyết định, AdaBoost, và mạng nơ ron với bài toán dự đoán thuê bao rời bỏ nhà mạng [9].

Cho đến nay, có một vài công trình áp dụng Rừng ngẫu nhiên để dự đoán thuê bao rời bỏ nhà mạng. Đầu tiên là công trình của Lariviere và Van den Poel, năm 2005 [20] áp dụng Rừng ngẫu nhiên để dự đoán thuê bao rời bỏ công ty dịch vụ tài chánh ở Châu Âu, tuy nhiên công trình này không xử lý vấn đề mất cân bằng dữ liệu có xảy ra trong ứng dụng. Gần đây, có một số công trình cố gắng cải tiến giải thuật Rừng ngẫu nhiên để xử lý vấn đề mất cân bằng dữ liệu trong bài toán dự đoán thuê bao rời bỏ nhà mạng. Effendy và các cộng sự, năm 2014, đề xuất một cách tiếp cận kết hợp lấy mẫu dữ liệu với Rừng ngẫu nhiên có gán trọng số (weighted random forest) để giải bài toán dự đoán thuê bao rời bỏ nhà mạng [21]. Paining và các cộng sự, năm 2018 so sánh hiệu quả của một số phương pháp lấy mẫu dữ liệu kết hợp với giải thuật Rừng ngẫu nhiên để phân lớp dữ liệu mất cân bằng [22]. Agusta và Adiwijava (2019) [23] ứng dụng một kỹ thuật lấy mẫu giảm dựa vào gom cụm cho mỗi lần tạo tập huấn luyện con trong giải thuật Rừng ngẫu nhiên để cải tiến giải thuật này cho bài toán dự đoán thuê bao rời bỏ nhà mạng. Devi và các cộng sự năm 2019 [4] đề xuất giải thuật Rừng ngẫu nhiên có điều chỉnh trọng số cho bài toán phát hiện gian lận thẻ tín dụng, cũng là bài toán có đặc tính mất cân bằng dữ liệu.

III. CÁCH TIẾP CẬN PHÂN LỚP ĐỀ XUẤT CHO BÀI TOÁN DỰ ĐOÁN THUÊ BAO RỜI BỎ NHÀ MẠNG CÓ ĐẶC TÍNH MẤT CÂN BẰNG DỮ LIỆU

Một trong những mục đích chính của nghiên cứu này là thử áp dụng giải thuật Rừng ngẫu nhiên có điều chỉnh hàm chi phí (CSWRF) được đề xuất bởi Devi và các cộng sự [4] cho bài toán phát hiện gian lận thẻ tín dụng sang một bối cảnh khác: bài toán dự đoán thuê bao rời bỏ nhà mạng. Giải thuật CSWRF sẽ được giải thích chi tiết trong mục này.

Giải thuật CSWRF bao gồm hai giai đoạn chính: giai đoạn huấn luyện và giai đoạn phân lớp (giai đoạn thử).

A. GIAI ĐOẠN HUẤN LUYỆN

Trong giai đoạn huấn luyện, giải thuật Rừng ngẫu nhiên chuẩn được cải tiến bằng cách thiết lập cách tính hàm chi phí dựa vào tỉ lệ phân lớp sai của các mẫu.

Tỉ lệ phân lớp sai của các mẫu được tính dựa vào số âm sai FN (false negative) và số dương sai FP (false positive) trong *ma trận đúng sai* (confusion matrix) của mỗi cây quyết định thành phần trong giai đoạn huấn luyện.

$$cost(T) = (\mathrm{FN}(T) + \mathrm{FP}(T)/|size(T)| \quad (1)$$

với FN(T) là số âm sai của cây T, FP(T) là số dương sai của cây T, và $|size(T)|$ là số lượng mẫu trong tập con huấn luyện của cây T.

Sau đó, một kỹ thuật dựa vào sai số sẽ gán trọng số cho mỗi cây quyết định. Sai số được tính dựa vào hai số hạng: *tổng sai số huấn luyện* (overall training error) và *tổng sai số huấn luyện lớp dương* (positive training error). Lưu ý là ở đây lớp dương ám chỉ lớp thiểu số.

Tổng sai số huấn luyện được tính dựa vào tổng số mẫu bị phân lớp sai bởi một cây quyết định. Tổng sai số huấn luyện lớp dương được tính dựa vào tổng số mẫu thuộc lớp dương bị phân lớp sai bởi một cây quyết định. Công thức tính của hai tổng sai số nêu trên được mô tả như sau:

$$err\_ov(x_i) = \begin{cases} 1: if\ y(x_i) \neq class(x_i) \\ 0: if\ y(x_i) = class(x_i) \end{cases} \quad (2)$$

$$err\_pos(x_m) = \begin{cases} 1: if\ y(x_m) \neq positive \\ 0: if\ y(x_m) = positive \end{cases} \quad (3)$$

Với x_i là một mẫu thuộc cây quyết định, $y(x_i)$ là lớp được gán cho mẫu x_i bởi cây quyết định, x_m là một mẫu thuộc lớp thiểu số của cây quyết định.

Trong trường hợp hai hay nhiều cây quyết định có cùng tổng sai số huấn luyện thì cây quyết định nào có tổng sai số lớp dương nhỏ hơn sẽ được gán trọng số cao hơn.

Hàm chi phí thể hiện khả năng phân lớp của cây quyết định. Cách tính trọng số của cây thể hiện độ tin cậy của cây. Cây quyết định có khả năng phân lớp cao nhất và độ tin cậy cao nhất được coi như cây quyết định tốt nhất.

Toàn bộ quá trình huấn luyện được mô tả trong Giải Thuật CSWRF_1.

Giải Thuật CSWRF _1

Input: S là tập dữ liệu, S_{min} là tập dữ liệu lớp thiểu số, S_{maj} là tập dữ liệu lớp đa số, n: số cây quyết định.

Output: Tập cây quyết định đã được gán trọng số: {$weight(T)$: T là một cây quyết định}

1. Tách tập S thành tập huấn luyện *train* và tập thử *test*.
2. Bằng cách lấy mẫu được phép trùng lặp để chia tập *train* thành n tập huấn luyện con, mỗi tập được ký hiệu là T.
3. **for** i := 1 **to** n **do**

(a) Tách tập T thành T_{maj} gồm các mẫu lớp đa số và T_{min} gồm các mẫu lớp thiểu số: $T_{maj} \subseteq S_{maj}$, $T_{min} \subseteq S_{\min}$.

(b) Huấn luyện bằng tập T để tạo ra một cây quyết định tương ứng.

(c) Tính tổng sai số huấn luyện của cây T dựa vào công thức (1) và công thức (2) như sau:

$$err\_ov(T) = \sum_{m=1}^{N} err\_ov(x_i) * cost(T) \quad (4)$$

N = tổng số mẫu của cây T.

(d) Tính tổng sai số huấn luyện lớp dương của cây T dựa vào công thức (1) và công thức (3) như sau:

$$err\_pos(T) = \sum_{m=1}^{p} err\_pos(x_m) * cost(T) \quad (5)$$

p = tổng số mẫu thuộc lớp thiểu số của cây T.

(e) Tính trọng số cho cây T_i với công thức $weight(T_i) = 1/err\_ov(T_i)$

Cây với tổng sai số huấn luyên nhỏ hơn sẽ có trọng số cao hơn.

(f) Trong trường hợp tổng sai số huấn luyện của hai hay nhiều cây bằng nhau, thì trọng số của chúng sẽ được tính bằng công thức sau:

$$weight(T_i) = 1/err\_pos(T_i)$$

Như vậy, một đặc tính nổi bật của giải thuật CSWRF là việc gán trọng số cho từng cây quyết định thành phần có tính đến tổng sai số huấn luyện lớp thiểu số của cây đó.

B. GIAI ĐOẠN PHÂN LỚP

Trong giai đoạn phân lớp, kết quả của mỗi mẫu trong tập thử được quyết định bằng cách dùng mỗi cây quyết định đã được huấn luyện để phân lớp mẫu thử và gom lại tất cả kết quả được xác định bởi tất cả các cây để đưa ra kết quả sau cùng.

Đối với giải thuật Rừng ngẫu nhiên thuần túy, kết quả (nhãn lớp) cuối cùng của một mẫu thử được xác định dựa vào phiếu bầu đa số (majority voting) của tất cả các kết quả từ mọi cây quyết định. Tuy nhiên, trong giải thuật CSWRF, kết quả cuối cùng được quyết định dựa vào trọng số của một cây quyết định đặc biệt. Cho một mẫu thử, x', quyết định cuối cùng của mẫu thử này, $F(x')$ được tính như sau:

$$F(x') = f(x') \text{ từ cây } T_i \mid weight(T_i) = \text{lớn nhất} \quad (6)$$

tức là kết quả từ *cây quyết định có trọng số lớn nhất* đối với mẫu thử sẽ được xem như kết quả cuối cùng của mẫu thử.

Toàn bộ quá trình phân lớp được mô tả trong Giải Thuật CSWRF_2.
Giải thuật CSWRF_2
Input: *test* là tập dữ liệu thử, W= {*weight*(T)} là tập hợp các trọng số của mọi cây quyết định T.
Ouput: tập thử mà mọi mẫu thử đã được phân lớp
1. Phân lớp mọi mẫu trong tập *test* bằng tất cả các cây quyết định thành phần
2. Với mỗi mẫu thử x' trong tập *test*
(a) với mỗi cây quyết định T_i trong rừng ngẫu nhiên
Ghi nhận nhãn lớp như là $y(x') = f(x')$ từ cây T_i
(b) xác định kết quả cuối cùng, $F(x')$ bằng cách dùng công thức (6).
Một đặc tính nổi bật đáng lưu ý của giải thuật CSWRF là việc huấn luyện mỗi cây quyết định thành phần trong rừng ngẫu nhiên có thể được tiến hành một cách độc lập.

IV. KẾT QUẢ THỰC NGHIỆM

Mục đích chính thứ hai của nghiên cứu này là khảo sát bằng thực nghiệm hiệu quả phân lớp của giải thuật CSWRF trong bài toán dự đoán thuê bao rời bỏ nhà mạng và so sánh hiệu quả của giải thuật CSWRF có sử dụng cách tiếp cận điều chỉnh hàm chi phí (cost-sensitive learning) với hiệu quả của sự kết hợp phương pháp lấy mẫu tăng SMOTE với giải thuật Rừng Ngẫu Nhiên thuần túy.

A. TẬP DỮ LIỆU THỰC NGHIỆM

Với bài toán dự đoán thuê bao rời bỏ nhà mạng, các thực nghiệm trong nghiên cứu này được thực hiện trên hai bộ dữ liệu mẫu: Telco Customer Churn và Cell2Cell, là hai bộ dữ liệu thường được dùng bởi cộng đồng nghiên cứu về bài toán dự đoán thuê bao rời bỏ nhà mạng.

Bảng 1. Những đặc điểm của hai tập dữ liệu mẫu

Tập dữ liệu	Telco Customer Churn	Cell2Cell
Nguồn	IBM	Duke University
Trang web	Kaggle	Kaggle
Số đặc trưng	19	58
Tổng số mẫu	7043	51047
Tổng số mẫu không rời bỏ	5174	36336
Tổng số mẫu rời bỏ	1869	14711
Các giá trị bị mất	không	có

Tập dữ liệu Telco Customer Churn dataset là từ công ty IBM. Bộ dữ liệu này được thu thập từ thuê bao của các nhà cung cấp dịch vụ viễn thông không dây và dữ liệu liên quan đến các cuộc gọi mà thuê bao thực hiện. Bộ dữ liệu được hoàn chỉnh bởi Kaggle [24].

Tập dữ liệu Cell2Cell được cung cấp bởi trung tâm dữ liệu Teradata Center cho bài toán quản lý quan hệ khách hàng của trường Đại học Duke. Tập dữ liệu này được thu thập từ các thuê bao của công ty viễn thông Cell2Cell. Cell2Cell là một trong những công ty lớn nhất của Hoa Kỳ về lãnh vực viễn thông không dây và tỉ lệ thuê bao rời bỏ hàng tháng là 4%.

Hai tập dữ liệu trên được tải xuống từ trang web của Kaggle [24]. Những đặc điểm của hai tập dữ liệu viễn thông này được mô tả ở Bảng 1.

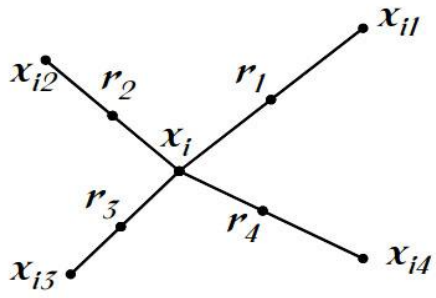
B. CÁC PHƯƠNG PHÁP PHÂN LỚP ĐỐI SÁNH

Chúng tôi thực nghiệm giải thuật Rừng ngẫu nhiên cải tiến CSWRF trên hai tập dữ liệu về thuê bao rời bỏ nhà mạng như đã mô tả ở mục IV.A. Để đánh giá hiệu quả phân lớp của giải thuật CSWRF, chúng tôi so sánh hiệu quả của giải thuật này với hai phương pháp so sánh: cây quyết định (ký hiệu là Decision Tree) và Rừng ngẫu nhiên thuần túy (ký hiệu là RF), là hai phương pháp chưa sử dụng kỹ thuật giải quyết vấn đề dữ liệu mất cân bằng. Để so sánh hiệu quả của hai cách tiếp cận nhằm xử lý vấn đề mất cân bằng dữ liệu, chúng tôi so sánh hiệu quả của cách tiếp cận điều chỉnh hàm chi phí, tức là giải thuật CSWRF, với cách tiếp cận lấy mẫu dữ liệu, tức là kết hợp giải thuật Rừng ngẫu nhiên thuần túy với phương pháp lấy mẫu tăng SMOTE (ký hiệu là RF + SMOTE).

Với phương pháp kết hợp Rừng ngẫu nhiên với SMOTE, trước khi áp dụng giải thuật Rừng ngẫu nhiên, kỹ thuật SMOTE được dùng để gia tăng số mẫu trong lớp thiểu số của tập dữ liệu. Ý tưởng chính của kỹ thuật SMOTE được mô tả như sau.

SMOTE làm giàu lớp thiểu số bằng cách chọn từng mẫu trong lớp thiểu số và sinh ra những mẫu mới dựa vào những đoạn thẳng nối từ mẫu đó đến bất kỳ hoặc tất cả *k* mẫu lân cận nhất của nó. Tương ứng với số lượng lấy mẫu tăng mong muốn, các lân cận từ *k* lân cận gần nhất sẽ được chọn một cách ngẫu nhiên. Một thí dụ đơn giản của SMOTE được minh họa trong Hình 2. Một mẫu thuộc lớp thiểu số x_i được chọn làm điểm căn bản để tạo ra nhiều điểm mẫu mới. Dựa vào một độ đo khoảng cách, nhiều lân cận gần nhất thuộc cùng lớp thiểu số (các điểm từ x_{i1} đến x_{i4}) được chọn từ tập huấn luyện. Sau đó, độ sai biệt giữa x_i và những điểm lân cận gần nhất của nó được tính và nhân với một giá trị ngẫu nhiên trong tầm [0, 1] để sinh ra những mẫu mới từ r_1 đến r_4. Những chi tiết rõ hơn về phương pháp SMOTE được mô tả trong bài báo [25].

Tất cả các thực nghiệm phân lớp trong nghiên cứu này được tiến hành dùng qui trình kiểm tra chéo 10-phần (10-fold cross validation).



Hình 2. *Minh họa cách tạo ra những điểm mẫu tổng hợp bằng phương pháp SMOTE [24] với k=4*

D. CÁC TIÊU CHÍ ĐÁNH GIÁ

Khi xử Iý dữ liệu mất cân bằng trong bài toán phân lớp, độ chính xác phân lớp toàn cục (overall classification accuracy) thường không phải là độ đo hiệu quả thích hợp. Với độ đo như vậy, một bộ phân lớp thông thường phân lớp mọi mẫu đều thuộc lớp đa số thì vẫn đạt được độ chính xác cao. Trong nghiên cứu này, chúng tôi sử dụng những độ đo như tỉ lệ âm đúng (true negative rate), tỉ lệ dương đúng (true positive rate), precision, độ truy hồi (recall), độ chính xác (accuracy), G-mean và F-measure (còn được gọi là F1-score) để đánh giá hiệu quả của các phương pháp phân lớp trên dữ liệu mất cân bằng. Vì trong ngữ cảnh phân lớp đặc biệt này, chúng tôi nhắm tới đạt được chất lượng tốt cho cả hai lớp, nên cần phải phối hợp nhiều độ đo riêng lẻ cho cả hai lớp: lớp thiểu số (lớp dương) và lớp đa số (lớp âm). Do đó, chúng tôi coi trọng hai độ đo: F-measure và G-mean. Tất cả các độ đo dựa vào ma trận đúng sai (confusion matrix) được mô tả ở Bảng 2. Các hàng trong ma trận diễn tả các lớp thực sự và các cột trong ma trận diễn cả các lớp được dự đoán. Dựa vào ma trận đúng sau ở Bảng 2, các độ đo đánh giá được tính bằng các công thức như sau:

$$\text{True Negative Rate } (\boldsymbol{Acc^-}) = \frac{\boldsymbol{TN}}{\boldsymbol{TN + FP}} \quad (7)$$

$$True\ Positive\ Rate\ (Acc^+) = \frac{TP}{TP + FN} \quad (8)$$

$$\text{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FP}} \quad (9)$$

$$\text{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FN}} \quad (10)$$

$$\text{Accuracy} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + TN + FP + FN}} \quad (11)$$

$$\text{F-measure} = \frac{\mathbf{2} \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (12)$$

$$\text{G-mean} = \sqrt{\text{Acc}^- \times \text{Acc}^+} \quad (13)$$

Bảng 2. *Ma trận đúng sai*

		Predicted class	
		P	N
Actual class	P	TP(True Positive)	FN(False
	N	FP(False Positive)	TN (True

E. KẾT QUẢ THỰC NGHIỆM

Trong phần thực nghiệm, số lượng cây quyết định trong giải thuật CSWRF được chọn là 100. Sự lựa chọn này là thông qua thực nghiệm: chúng tôi thử nghiệm giải thuật CSWRF với nhiều số lượng cây khác nhau và tìm thấy giá trị 100 là phù hợp nhất. Theo các công trình đi trước, số lượng cây không ảnh hưởng nhiều đến chất lượng phân lớp của Rừng ngẫu nhiên mà có ảnh hưởng đến thời gian thực thi của giải thuật này.

Chúng tôi hiện thực Cây quyết định và Rừng ngẫu nhiên dựa vào thư viện *scikit-learn*. Chúng tôi hiện thực SMOTE với công cụ *imbalanced-learn* [26] và hiện thực giải thuật CSWRF với ngôn ngữ lập trình Python.

Trong nghiên cứu này, chúng tôi sử dụng *F-measure* (*positive class*) làm độ đo đánh giá chính. Ngoài ra, *F-measure* (*negative class*), *G-mean* và *Accuracy* được dùng như là những độ đo đánh giá phụ.

Bảng 3 và Bảng 4 trình bày những độ đo kết quả (trung bình) của các phương pháp phân lớp đối sánh khi xử lý vấn đề dữ liệu mất cân bằng trên hai tập dữ liệu mẫu Telco và Cell2Cell. Những kết quả thực nghiệm được in đậm là những kết quả của phương pháp đạt hiệu quả trung bình cao nhất.

Bảng 3. *Kết quả về các độ đo trên tập dữ liệu TELCO*

	Acc	**G-Mean**	**F-measure (Neg)**	**F-measure (Pos)**
Decision Tree	0.7243	0.6281	0.8117	0.4841
RF	0.7894	0.6615	0.8624	0.5509
RF +SMOTE	0.7828	0.6729	0.8560	0.5582
CSWRF	**0.8028**	**0.7322**	**0.8663**	**0.6240**

Bảng 4. *Kết quả về các độ đo trên tập dữ liệu CELL2CELL*

	Acc	**G-Mean**	**F-measure (Neg)**	**F-measure (Pos)**
Decision Tree	0.6200	0.5511	0.7300	0.3534
RF	0.7202	0.2989	0.8322	0.1593
RF +SMOTE	0.7199	0.3207	0.8312	0.1778
CSWRF	**0.7659**	**0.7044**	**0.8353**	**0.5957**

Từ những kết quả thực nghiệm ở Bảng 3 và Bảng 4, chúng ta có thể thấy:

- Nhờ vào sự điều chỉnh hàm chi phí, giải thuật phân lớp Rừng ngẫu nhiên cải tiến (CSWRF) đã đem lại hiệu quả phân lớp tốt hơn hai giải thuật truyền thống chưa quan tâm đến vấn đề mất cân bằng dữ liệu: Cây quyết định và Rừng ngẫu nhiên trong bài toán dự đoán thuê bao rời bỏ nhà mạng. Kết quả thực nghiệm còn cho thấy giải thuật CSWRF không chỉ hiệu quả khi xử lý bài toán phát hiện gian lận thẻ tín dụng mà còn hiệu quả khi xử lý bài toán dự đoán thuê bao rời bỏ nhà mạng.
- Hiệu quả phân lớp của giải thuật CSWRF dựa vào sự điều chỉnh hàm chi phí khi phân lớp dữ liệu mất cân bằng thì tốt hơn hiệu quả phân lớp của phương pháp dựa vào lấy mẫu dữ liệu, tức là kết hợp lấy mẫu giảm SMOTE trước khi áp dụng giải thuật Rừng ngẫu nhiên.

Thời gian huấn luyện và thời gian phân lớp (tính bằng mili-giây) của bốn phương pháp đối sánh trên hai bộ dữ liệu mẫu được cho ở Bảng 5 và Bảng 6.

Từ Bảng 5, chúng ta có thể thấy thời gian huấn luyện của giải thuật CSWRF cao hơn đôi chút so với giải thuật RF trên 2 tập dữ liệu mẫu và thời gian huấn luyện của phương pháp RF+SMOTE cao hơn nhiều so với CSWRF. Thời gian huấn luyện của RF+SMOTE cao hơn nhiều so với CSWRF là do phương pháp lấy mẫu tăng SMOTE gây ra một chi phí tính toán đáng kể cho công tác làm tăng số lượng mẫu của lớp thiểu số. Thời gian huấn luyện của cả

ba phương pháp RF, RF+SMOTE và CSWRF đều cao hơn Cây quyết định là vì Rừng ngẫu nhiên bao gồm 100 cây quyết định.

Bảng 5. *Thời gian huấn luyện của 4 phương pháp trên hai tập dữ liệu*

	Telco	**Cell2Cell**
Decision Tree	348	12250
RF	6720	146786
RF +SMOTE	9000	191652
CSWRF	6780	152216

Bảng 6. *Thời gian phân lớp của 4 phương pháp trên hai bộ dữ liệu*

	Telco	**Cell2Cell**
Decision Tree	1	5
RF	32.1	261
RF +SMOTE	31.3	272
CSWRF	1	6

Từ Bảng 6, chúng ta có thể thấy thời gian phân lớp của CSWRF thấp hơn nhiều so với RF trên 2 tập dữ liệu mẫu và thời gian phân lớp của RF+SMOTE cao hơn rất nhiều so với CSWRF. Thời gian phân lớp của CSWRF thấp hơn nhiều so với RF là vì việc đánh trọng số mỗi cây của CSWRF giúp ra quyết định lớp cho mẫu thử nhanh hơn cách dựa vào nguyên tắc đa số phiếu của RF. Thời gian phân lớp của RF+SMOTE cao hơn rất nhiều so với CSWRF là vì khi phân lớp RF+SMOTE làm việc trên bộ dữ liệu với số lượng mẫu của lớp thiểu số lớn hơn rất nhiều so với số lượng mẫu của lớp thiểu số có trong CSWRF.

V. KẾT LUẬN

Giải thuật Rừng ngẫu nhiên cải tiến CSWRF, được đề xuất bởi Devi và các cộng sự năm 2019 [4] để xử lý vấn đề mất cân bằng dữ liệu thông qua điều chỉnh hàm chi phí cho bài toán phát hiện gian lận thẻ tín dụng. Trong nghiên cứu này, chúng tôi ứng dụng giải thuật CSWRF vào bài toán dự đoán thuê bao rời bỏ nhà mạng và so sánh hiệu quả của giải thuật CSWRF với một phương pháp xử lý vấn đề mất cân bằng dữ liệu thông qua kỹ thuật lấy mẫu dữ liệu: Rừng ngẫu nhiên kết hợp SMOTE. Kết quả thực nghiệm trên hai tập dữ liệu mẫu cho thấy giải thuật CSWRF có xử lý vấn đề mất cân bằng dữ liệu phân lớp hiệu quả hơn hai phương pháp truyền thống không xử lý vấn đề mất cân bằng dữ liệu: Cây quyết định và Rừng ngẫu nhiên. Ngoài ra kết quả thực nghiệm còn cho thấy giải thuật CSWRF dựa vào sự điều chỉnh hàm chi phí hiệu quả hơn phương pháp dựa vào kỹ thuật lấy mẫu dữ liệu như Rừng ngẫu nhiên kết hợp SMOTE. Thời gian huấn luyện của giải thuật CSWRF thấp hơn so với phương pháp Rừng ngẫu nhiên kết hợp SMOTE cũng cho thấy tính khả dụng cao của CSWRFR trong những ứng dụng thực tế. Tóm lại, một kết luận quan trọng rút ra từ công trình này là giải thuật CSWRF không chỉ hiệu quả đối với bài toán phát hiện gian lận thẻ tín dụng mà còn hiệu quả đối với bài toán dự đoán thuê bao rời bỏ nhà mạng .

Trong tương lai, chúng tôi dự định mở rộng giải thuật này bằng cách tích hợp thêm kỹ thuật lựa chọn đặc trưng để giải thuật có thể làm việc với những tập dữ liệu số chiều nhiều [27]. Tiếp theo, chúng tôi dự tính sẽ so sánh hiệu quả của giải thuật CSWRF với giải thuật XGBoost kết hợp Focal loss trong bài toán phát hiện thuê bao rời bỏ nhà mạng. Ngoài ra, chúng tôi dự định hiện thực giải thuật CSWRF trên môi trường GPU (Graphics Processing Unit) để tăng tốc quá trình tính toán của giải thuật.

TÀI LIỆU THAM KHẢO

[1] V. Umayaparvathi, K. Iyakutti, "A survery on customer churn prediction in telecom industry: datasets, methods and metrics", *Int. Research Journal of Engineering and Technology*, vol. 3, no. 4, pp. 1065-1070, 2016.

[2] V. Lopez, A. Fernandez, S. Garcia, V. Palade, F. Herreta, "An insight into classification with imbalance data: Empirical results and current trends on using data intrisic characteristics", *Information Science*, vol. 20, pp. 113-141, 2013.

[3] N. N. Nam, D. T. Anh, "Comparison of Two Main Approaches for Handling Imbalanced Data in Churn Prediction Problem," *Journal of Advances in Information Technology*, vol. 12, No. 1, pp. 29-35, Feb. 2021

[4] D. Devi, S.K. Biswas, B. Purkayastha, "A Cost-sensitive weighted Random Forest Technique for Credit Card Fraud Detection," *Proc. of Int. Conf. on Computing and Networking Technology* (ICCNT), Kanpur, India, 6-8 July, 2019.

[5] A. Sharma, P. K. Panigrahi, "A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services," *International Journal of Computer Applications,* vol. 27, 2013.

[6] Y. Zhao, B. Li, X. Li, W. Liu, S. Ren, "Customer Churn Prediction Using Improved One-Class Support Vector Machine," *Proc. of Advanced Data Mining and Applications* (ADMA), Berlin, LNCS 3584, Springer, pp. 300-306, 2005.

[7] Y. Zhang, J. Qi, H. Shu, J. Cao, "A hybrid KNN-LR classifier and its application in customer churn prediction," *Proc. of IEEE International Conference on Systems, Man and Cybernetics*, Montréal, Canada, 7-10 Oct, 2007.

[8] N. Lu, H. Lin, J. Lu, "A customer churn prediction model in telecom industry using boosting", *IEEE Transactions on Industrial Informatics*, vol. 10, no. 2, pp. 1659 – 1665, 2014.

[9] G. Esteves and J. Mendes-Moreira, "Churn perdiction in the telecom business," *Proc. of 11th International Conference on Digital Information Management* (ICDIM), 19-21 Sep., Porto, Portugal, 2016.

[10] G. Li, X. Deng , "Customer Churn Prediction of China Telecom Based on Cluster Analysis and Decision Tree Algorithm," *Proc. of Emerging Research in Artif. Intel. and Computational Intelligence*, Chengdu, China, 2012.

[11] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research,* vol. 16, pp. 321–357, 2002.

[12] M. A. Tahir, J. Kittler, K. Mikolajczyk, F. Yan, "A multiple expert approach to class imbalance problem using inverse random undersampling", *Proc. of Int. Workshop on Multiple Classifier Systems*, pp. 82-91, 2009.

[13] W. C. Lin, C.F. Tsai, Y. H. Hu, J. S. Jhang, "Clustering-based undersampling in class imbalanced data", *Information Sciences*, vol. 409-410, pp. 17-26, 2012.

[14] G. Haixang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, "Learning from class-imbalanced data: Reiew of methods and applications", *Expert Systems With Application*, vol. 73, May, pp. 220-239, 2016.

[15] G. Weiss, "Mining with rarity: a unifying framework.", *Journal of Artificial Intelligence Research*, vol. 19, pp. 315-334, 2004.

[16] T. Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, "Focal loss for dense object selection", *Proc. of IEEE Int. Conf. on Computer Vision and Applications*, pp. 1243-1248, 2018.

[17] C. Wang, C. Deng, S. Wang, "Imbalance-XGBoost: Leveraging Weighted and Focal Losses for Binary Label-Imbalanced Classification with XGBoost," *Pattern Recognition Letter*, vol. 136, pp. 190-197, 2020.

[18] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system", Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, 2016.

[19] L. Breiman, "Random Forests", *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[20] B. Larivière, D.V.D. Poel, "Predicting customer retention and profitability by using random forests and regression forests techniques," *Expert Systems with Applications,* vol. 29, no. 2, pp. 472-484, 2005.

[21] V. Effendy, Adiwijaya, Z. K. A. Baizal, "Handling imbalanced data in customer churn prediction using combined sampling and weighted random forest," *Proc. of 2nd International Conference on Information and Communication Technology* (*ICoICT*), pp. 325-330, 2014.

[22] M. P. Paing, C. Pintavirooj, K. Hamamoto, "Comparision of sampling methods for imbalanced data classification in Random Forest", *Proc. of Biomedical Engineering International Conference* (BMEiCON), Chiang Mai, Thailand, 21-24 Nov., 2018.

[23] Z. P. Agusta, Adiwijaya, "Modified balanced random forest for improving imbalanced data prediction", *International Journal of Advances in Intelligent Informatics*, vol. 5, no. 1, pp. 58-65, March 2019.

[24] Kaggle web page, https://www.kaggle.com/ (accessed in 2022).

[25] A. Fernandez, S. Garcia, F. Herera, N. V. Chawla, "SMOTE for learning from imbalance data: progress and challenges, making the 15-year aniversary", *Journal of Artif. Intelligence Research,* vol. 61, pp. 863–905, 2018.

[26] G. Lemaitre, N. Nogueira, C. K. Aridas, "Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning", *Journal of Machine Learning Research*, vol. 18, pp. 1-5, 2017.

[27] A. Hanif, N. Azhar, "Resolving class imbalance and feature selection in customer churn dataset", *Proc. of Int. Conf. on Frontiers of Information Technology,* 18-20 Dec., Islamabad, Pakistan, pp. 82-86, 2017.

CLASSIFICATION WITH IMBALANCED DATA IN CUSTOMER CHURN PREDICTION BASED ON IMPROVE RANDOM FOREST

Dinh Minh Hoa, Duong Tuan Anh

ABSTRACT— In the telecommunication industry, customer churn is of great interest since this problem can affect the company's profit. However, the imbalanced data in customer churn prediction caused difficulties in developing a good prediction model for solving this problem. In this work, we proposed a random forest-based approach for classification with imbalanced data in telecom customer churn prediction. This approach utilizes the cost-sensitive weighted random forest (CSWRF), which was proposed for credit card fraud detection prediction. We compare the performance of CSWRF against one data resampling method: random forest combined with data sampling SMOTE (Synthetic Minority Oversampling Technique). Our experiments on two benchmark datasets reveal that for churn prediction in telecom which is an imbalanced data problem, the classification performance of CSWRF method is better than that of SMOTE combined with random forest..

Keywords— Imbalanced data, customer churn predicion, random forest, cost-sensitive learning, SMOTE.

Đinh Minh Hòa tốt nghiệp cử nhân công nghệ thông tin tại Trường Đại Học Ngoại Ngữ-Tin Học Tp. Hồ Chí Minh năm 2015 và tốt nghiệp thạc sĩ ngành công nghệ thông tin tại cùng trường đại học nêu trên năm 2022. Anh là lập trình viên tại công ty TNHH VinaHost Việt Nam từ 2015 đến 2018 và là trưởng phòng kỹ thuật công ty 7Host Việt Nam từ 2018 đến 2022. Từ năm 2023, anh là giảng viên khoa Công Nghệ Thông Tin của Trường Đại học Ngoại ngữ -Tin học Tp. Hồ Chí Minh.

Dương Tuấn Anh tốt nghiệp tiến sĩ ngành khoa học máy tính tại Viện Công Nghệ Á Châu (Asian Institute of Technology), Bangkok, Thái Lan, năm 1998 và đó cũng là nơi mà ông tốt nghiệp thạc sĩ cùng chuyên ngành. Ông đã là Phó Giáo sư tại Khoa Khoa Học và Kỹ thuật máy tính, trường Đại học Bách Khoa, ĐHQG Tp. Hồ Chí Minh từ năm 2007. Hiện nay, ông là giảng viên khoa Công nghệ thông tin, trường Đại học Ngoại ngữ-Tin học Tp. Hồ Chí Minh. Lãnh vực nghiên cứu chính của ông là metaheuristics, học máy và khai phá dữ liệu chuỗi thời gian. Ông là đồng tác giả của trên 120 bài báo khoa học.

ô