PHƯƠNG PHÁP MỚI XỬ LÝ DỮ LIỆU MẤT CÂN BẰNG NÂNG CAO HIỆU QUẢ DỰ ĐOÁN KHÁCH HÀNG RỜI BỎ DỊCH VỤ

TS. Nguyễn Hữu Xuân Trường* - Ths. Nguyễn Văn Tuấn*
Ths. Lê Xuân Đoàn* - TS. Đặng Xuân Thọ**

Dự đoán khách hàng rời bỏ dịch vụ là một bài toán phân lớp trong khai phá dữ liệu, sử dụng các mô hình phân lớp để dự đoán khách hàng có khả năng rời bỏ dịch vụ hay không. Đây là một trong những bài toán khó và có ý nghĩa quan trọng đối với các doanh nghiệp. Bài toán này đặc biệt khó bởi dữ liệu thường gặp vấn đề mất cân bằng khi số lượng khách hàng rời dịch vụ chỉ chiếm một tỉ lệ nhỏ trong tổng số. Do đó, bài toán dự đoán khách hàng rời dịch vụ trở nên khó khăn và thách thức hơn, cần có hướng tiếp cận mới để giải quyết. Một số phương pháp phổ biến giải quyết vấn đề này như SMOTE, Borderline-SMOTE, Safe-level SMOTE mặc dù đã đạt được những kết quả tích cực, nhưng một số trường hợp lại không đạt được kết quả mong đợi. Trong bài báo này, chúng tôi sẽ trình bày về ứng dụng của phương pháp phân lớp dữ liệu mất cân bằng trong giải quyết bài toán dự đoán khách hàng rời bỏ dịch vụ và đưa ra phương pháp cải tiến mới để nâng cao hiệu quả dự đoán.

• Từ khóa: *dự đoán khách hàng rời bỏ dịch vụ, dữ liệu mất cân bằng, phân lớp, SMOTE.*

customer churn prediction is a classification problem in data mining, which uses classification models to predict whether customers are likely to leave a service or not. It is one of the difficult and important problems for businesses. This problem is especially difficult because the data often has an imbalanced problem when the number of customers leaving the service only accounts for a small percentage of the total. Therefore, the problem of predicting customer leaving the service becomes more difficult and challenging, requiring a new approach to solve. Some popular methods of solving this problem such as SMOTE, Borderline-SMOTE, Safe-level SMOTE, although achieving positive results, in some cases did not achieve the expected results. In this paper, we will present the application of the imbalanced data classification method in solving the problem of predicting customer churn and offer annew method to improve prediction efficiency.

• Keywords: *customer churn, imbalanced data, classification, SMOTE.*

Ngày nhận bài: 25/122021
Ngày gửi phản biện: 26/12/2021
Ngày nhận kết quả phản biện: 30/12/2021
Ngày chấp nhận đăng: 30/01/2022

1. Giới thiệu

Bài toán dự đoán khách hàng rời bỏ giao dịch vụ là dự đoán liệu khách hàng sẽ không còn mua sản phẩm hoặc dịch vụ của mình trong một khoảng thời gian nhất định nữa hay không. Trong thực tế thì bài toán này thường gặp trong các tổ chức kinh doanh dịch vụ như ngân hàng, bảo hiểm, viễn thông... Đây là một bài toán quan trọng đối với bất kỳ một tổ chức nào vì nếu có thể dự đoán được sớm việc khách hàng không tiếp tục sử dụng dịch vụ, tổ chức có thể đưa ra được các phương án để giữ chân khách hàng. Việc chú trọng đến tập khách hàng (có khả năng) rời bỏ dịch vụ luôn được Ban lãnh đạo của các Tổ chức quan tâm vì nhiều lý do. Bởi giữ chân khách hàng sẽ giúp tăng uy tín thương hiệu, tăng doanh thu. Bên cạnh đó, chi phí đầu tư mỗi khách hàng mới nhiều gấp nhiều lần khách hàng cũ và việc tìm kiếm khách hàng mới cũng sẽ bị ảnh hưởng bởi việc khách cũ rời bỏ dịch vụ. Nhận thức được những điều đó, các tổ chức luôn cố gắng níu kéo từng khách hàng một, tìm biện pháp để kịp thời giữ chân khách hàng có nguy cơ rời bỏ dịch vụ của mình (Duyen, 2017).

Dự đoán khách hàng rời bỏ dịch vụ nói riêng và các bài toán dự báo trong kinh tế, tài chính nói chung đều đã có những nghiên cứu phân

** Học viện Chính sách và Phát triển ** Đại học Sư phạm Hà Nội*

tích dữ liệu để giải quyết từ lâu nhưng đến nay vẫn luôn được quan tâm đặc biệt bởi tầm quan trọng của nó. Với sự phát triển của khoa học công nghệ và sự bùng nổ dữ liệu hiện nay xuất hiện các kho dữ liệu khổng lồ (Big Data) thì các phương pháp phân tích dữ liệu truyền thống đòi hỏi những yêu cầu điều tra phức tạp và tốn kém về mặt thời gian. Do đó, xu thế hiện nay để giải quyết hiệu quả hơn các bài toán này là sử dụng các kỹ thuật của khai phá dữ liệu và các thuật toán học máy (Nguyễn Ngọc Tuân, 2016), (H. Ali, 2019).

Bài toán dự đoán khách hàng rời bỏ dịch vụ là thuộc dạng phân lớp trong khai phá dữ liệu và có đặc thù dữ liệu thường là mất cân bằng khi số lượng đa số (không rời dịch vụ) có thể sẽ lớn hơn rất nhiều so với số lượng lớp thiểu số (có rời dịch vụ), điều này làm cho các thuật toán phân lớp gặp rất nhiều khó khăn,do đó cần có hướng tiếp cận riêng để giải quyết. (Yanmin , 2009). Mặc dù hiện nay đã có một số phương pháp, thuật toán được đề xuất cho mô hình phân lớp dự đoán khách hàng rời bỏ dịch vụ và đã thu được những kết quả nhất định trong một số trường hợp riêng, tuy nhiên vấn đề này có thể được làm tốt hơn nữa để nâng cao hiệu quả dự đoán.

2. Phương pháp phân lớp dữ liệu mất cân bằng

2.1. Phân lớp dữ liệu

Phân lớp (classification) là một kỹ thuật quan trọng trong khai phá dữ liệu, mục đích là gán (dự đoán) nhãn của một phần tử dữ liệu mới (chưa biết nhãn) từ những thuộc tính của phần tử dữ liệu đó. Tập các giá trị nhãn lớp ở đây là hữu hạn, và nếu chỉ có 2 giá trị thì được gọi là phân lớp nhị phân. Ví dụ điển hình của phân lớp dữ liệu như việc phân loại email mới gửi đến là thư rác hay không, nếu là thư rác thì email sẽ được gán nhãn Spam và chuyển vào thư mục spam, còn nếu không thì sẽ được gán nhãn Non-spam và chuyển vào thư mục inbox.

Quá trình phân lớp gồm hai giai đoạn: xây dựng mô hình (learning) và sử dụng mô hình (classification).Giai đoạn xây dựng mô hình là việc học dữ liệu từ một tập dữ liệu huấn luyện (training set) đã biết trước nhãn bằng các thuật toán học máy (machine learning) để tạo ra một mô hình (model) có khả năng dự đoán nhãn lớp cho dữ liệu mới. Tuỳ theo thuật toán học máy được sử dụng thì có những mô hình phân lớp khác nhau, chẳng hạn như: cây quyết định (Decision Tree), k - láng giềng gần nhất (k - Nearest Neighboor), máy véc tơ hỗ trợ (Support Vector Machine), Naïve Bayes, rừng ngẫu nhiên (Random Forest)…

Sau khi xây dựng được mô hình phân lớp ở giai đoạn huấn luyện thì sẽ sử dụng mô hình để phân lớp dữ liệu mới nếu hiệu quả phân lớp chấp nhận được. Để đánh giá mô hình phân lớp là chấp nhận được hay không, ta sử dụng một bộ dữ liệu kiểm tra độc lập với bộ dữ liệu huấn luyện rồi từ đó xác định xem có bao nhiêu phần tử dữ liệu được phân lớp đúng và bao nhiêu phân tử dữ liệu bị phân lớp sai. Một số độ đo đánh giá hiệu quả phân lớp phổ biến là: *Accuracy*, *F-score*, *Sensitivity* (*Recall*), *Specificity*, *G-mean*… (Yanmin Sun, 2009).

2.2. Phân lớp dữ liệu mất cân bằng

Dữ liệu mất cân bằng là dữ liệu có sự chênh lệch lớn về số lượng phần tử giữa các lớp dự đoán (H. Ali, 2019), nghĩa là số lượng các phần tử đại diện cho một lớp lớn hơn rất nhiều so với các lớp khác, chẳng hạn như tỷ lệ khách hàng không rời bỏ dịch vụ thường là cao hơn rất nhiều so với khách hàng rời bỏ dịch vụ. Hoặc trong việc phát hiện bệnh nhân ung thư thì tỷ lệ bệnh nhân không bị ung thư là cao hơn rất nhiều so với các bệnh nhân bị ung thư... Đối với trường hợp hai lớp (chỉ có 2 trường hợp của lớp dự đoán, chẳng hạn như việc xác định giới tính là Nam/Nữ, hay xác định khả năng khách hàng là có rời bỏ dịch vụ/ không rời bỏ dịch vụ...) thì tỷ lệ này có thể là 1:5, 1:10, 1:100… Lớp chiếm số đông phần tử gọi là lớp đa số (negative), ngược lại lớp có ít phần tử gọi là lớp thiểu số (positive). Khi tiến hành khai phá dữ liệu trên các dữ liệu mất cân bằng thì các thuật toán thường đạt độ chính xác cao với lớp đa số nhưng với lớp thiểu số thì ngược lại.

Trong thực tế, vấn đề mất cân bằng dữ liệu đối với các bộ dữ liệu của bài toán dự đoán khách hàng rời bỏ dịch vụ là phổ biến. Để giải quyết hiệu quả cho bài toán này thì có thể sử dụng các phương pháp phân lớp của khai phá dữ liệu, trong đó có hai hướng tiếp cận chính được tập trung nghiên cứu cho việc nâng cao hiệu quả dự đoán là: hướng tiếp cận ở mức độ thuật toán và hướng tiếp cận ở mức độ dữ liệu (Yanmin Sun, 2009), (H.Ali, 2019).

Hướng tiếp cận ở mức độ thuật toán: Tập trung vào việc điều chỉnh, cải tiến các thuật toán phân lớp chuẩn (như cây quyết định, Naïve Bayes, máy véctơ hỗ trợ SVM, k láng giềng gần nhất KNN...) sao cho phù hợp với dữ liệu mất cân bằng, chẳng hạn như tăng cường học cho lớp thiểu số. Hướng tiếp cận này thường là phức tạp hơn so với hướng tiếp cận ở mức độ dữ liệu và yêu cầu cần phải hiểu rõ về thuật toán phân lớp cần cải tiến.

Hướng tiếp cận ở mức độ dữ liệu: Bao gồm các phương pháp điều chỉnh để giảm sự mất cân bằng dữ liệu bằng cách tăng số lượng phần tử lớp thiểu số (sinh thêm các phần tử thuộc lớp thiểu số một cách ngẫu nhiên, hoặc có chọn lọc, hoặc sinh thêm phần tử nhân tạo), giảm số lượng phần tử lớp đa số hoặc kết hợp cả hai phương pháp. Cả ba phương pháp trên đều hướng đến mục tiêu cân bằng phân bố dữ liệu. Ưu điểm của các phương pháp tiếp cận mức độ dữ liệu là sự linh hoạt, dữ liệu có thể sử dụng để huấn luyện các bộ phân loại khác nhau. Hướng tiếp cận này được tập trung nghiên cứu nhiều hơn và một số phương pháp tiêu biểu có thể kể tới là:

• ROS (Random Over-sampling)

• RUS (Random Under-sampling)

• SMOTE (Synthetic Minority Over-sampling Technique)

• BOS (Boderline SMOTE): sinh thêm phần tử nhân tạo dựa trên đường biên

• SLS (Safe-level SMOTE): sinh thêm phần tử nhân tạo dựa trên mức an toàn

• Một số phương pháp khác: Tomek Link, ADASYN…

Hình 1. Phân lớp dữ liệu mất cân bằng theo hướng tiếp cận mức độ dữ liệu

Một số tác giả cũng đã sử dụng phương pháp phân lớp dữ liệu mất cân bằng để giải quyết bài toán dự đoán khách hàng rời bỏ dịch vụ theo các cách tiếp cận khác nhau. Nhóm tác giả trong (Aamer, 2017) đã kết hợp phương pháp lựa chọn thuộc tính và xử lý dữ liệu mất cân bằng, trong khi đó, Annisa Aditsania và các cộng sự trong (Annisa, 2017) sử dụng phương pháp lấy mẫu tổng hợp thích ứng (ADASYN - một biến thể của SMOTE) và thuật toán lan truyền ngược để xử lý mất cân bằng dữ liệu trong bài toán dự đoán khách hàng rời bỏ dịch vụ. Ngoài ra, các tác giả trong (Uma, 2018) đã đề xuất phương pháp mới SOS-BUS kết hợp tăng số phần tử nhân tạo SMOTE với kỹ thuật giảm số phần tử của họ đề xuất… Về cơ bản là các nghiên cứu tập trung chủ yếu vào việc tiếp cận xử lý mất cân bằng ở cấp độ dữ liệu, đề xuất một số cải tiến kết hợp với thuật toán SMOTE để dự báo khách hàng rời bỏ dịch vụ, tuy nhiên dữ liệu sử dụng mang tính đặc thù và không công khai.

Tại Việt Nam, một số tác giả như (Nguyễn Ngọc Tuân, 2016), (Kien Vu, 2018) đã nghiên cứu việc áp dụng kỹ thuật khai phá dữ liệu để giải quyết bài toán dự đoán khách hàng rời bỏ dịch vụ trong lĩnh vực kinh doanh viễn thông. Đặc biệt, nhóm tác giả của FTP Telecom (Duyen, 2017) và các đồng nghiệp cũng đã có ứng dụng phân lớp dữ liệu mất cân bằng và phương pháp lựa chọn thuộc tính để giải quyết bài toán dự đoán khách hàng rời bỏ dịch vụ Internet tại tổ chức của mình.

3. Phương pháp đề xuất nâng cao hiệu quả dự đoán khách hàng rời bỏ dịch vụ

Mặc dù các phương pháp phân lớp dữ liệu mất cân bằng phổ biến hiện nay như SMOTE và các biến thể của nó đã có những cải tiến mô hình phân lớp bằng cách sinh thêm các phần tử nhân tạo

theo những cách khác nhau, tuy nhiên qua khảo sát dữ liệu của một số bài toán dự đoán khách hàng rời bỏ dịch vụ và thực nghiệm thì chúng tôi nhận thấy rằng các phương pháp này còn tồn tại một số hạn chế như:

• Phần tử thiểu số nhân tạo được sinh quá xa với phần tử thiểu số thực sự, và có thể sẽ nằm trong vùng có nhiều phần tử đa số nên sẽ gây nhiễu (giảm hiệu quả phân lớp).

• Những phần tử đa số nằm sâu trong vùng có nhiều phần tử thiểu số là những phần tử có ảnh hưởng nhiễu lớn nhưng các phương pháp hiện tại không xử lý nhiễu với những phần tử này.

• Phương pháp Borderline SMOTE thì chỉ áp dụng cho các trường hợp dữ liệu được phân bố bởi đường biên rõ ràng, hay phương pháp Safe-level SMOTE thì định nghĩa tỷ lệ an toàn dựa trên số phần tử láng giềng có thể sẽ bị nhiễu nếu phân bố dữ liệu không đồng đều về khoảng cách nên khi áp dụng với bài toán khách hàng rời bỏ dịch vụ thì hiệu quả không cao…

Do đó, chúng tôi đề xuất một phương pháp mới để nâng cao hiệu quả cho mô hình dự đoán khách hàng rời bỏ dịch vụ. Phương pháp này cũng giống như SMOTE là xử lý dữ liệu mất cân bằng trước khi huấn luyện mô hình, các bước thực hiện như sau:

• Lấy giá trị khoảng cách R_1 và R_2 đủ nhỏ và tham số $T \geq 1$.

• Với mỗi phần tử đa số X, nếu $T*n > m$, trong đó lần lượt là số phần tử đa số, số phần tử thiểu số có khoảng cách tới X nhỏ hơn R_1 thì đổi nhãn của X thành nhãn của lớp thiểu số.

• Với mỗi phần tử thiểu số thực sự P, sinh thêm ngẫu nhiên k phần tử thiểu số nhân tạo có khoảng cách tới P nhỏ hơn R_2.

Để minh chứng tính hiệu quả của phương pháp đề xuất, chúng tôi tiến hành thực nghiệm trên bộ dữ liệu Customer Churn của Kaggle và so sánh kết quả với các phương pháp xử lý dữ liệu mất cân bằng khác, bao gồm: SMOTE, Boderline SMOTE (BLS), Safe-level SMOTE (SLS). Bộ dữ liệu này có kích thước gồm 4250 bản ghi, trong đó tỷ lệ mất cân bằng dữ liệu là 1 : 6,1. Các thuật toán phân lớp sử dụng trong thực nghiệm là: rừng ngẫu nhiên (RF), máy véc tơ hỗ trợ (SVM), k láng giềng gần nhất (KNN).

Hình 2. Các bước tiến hành thực nghiệm

Chúng tôi sử dụng các độ đo *Sensitivity (Recall), Specificity, G-mean* được tính từ ma trận nhầm lẫn như sau:

Bảng 1. Ma trận nhầm lẫn

	Dự đoán là Positive	Dự đoán là Negative
Thực tế là Positive	TP	FN
Thực tế là Negative	FP	TN

Trong đó các hàng của ma trận là nhãn lớp thực tế, các cột là nhãn lớp dự đoán và:

• TN: số lượng phần tử lớp đa số được phân loại chính xác.

• FN: số lượng phần tử lớp thiểu số bị phân loại nhầm là phần tử lớp đa số.

• TP: số lượng phần tử lớp thiểu số được phân loại chính xác.

• FP: số lượng phần tử lớp đa số bị phân loại nhầm là phần tử lớp thiểu số.

• $Sensitivity\ \frac{TP}{TP+FN}$: tỷ lệ phát hiện ra các phần tử thiểu số thực sự.

• $Specificity = \frac{TN}{TN+FP}$: tỷ lệ phát hiện ra các phần tử đa số thực sự.

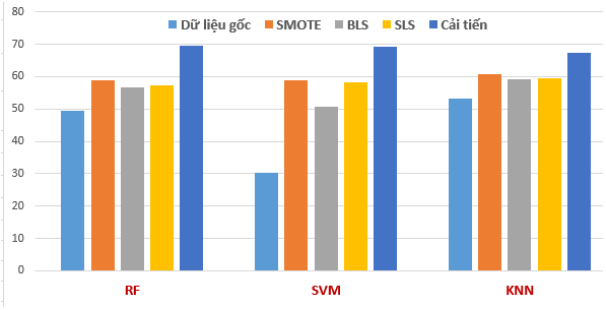
• $G-mean = \sqrt{Sensitivity\ *\ Specificity}$

Kết quả thực nghiệm được thống kê như sau:

Bảng 2. Tổng hợp kết quả thực nghiệm

Phương pháp	RF			SVM			KNN		
	Sen	Spe	G-mean	Sen	Spe	G-mean	Sen	Spe	G-mean
Dữ liệu gốc	25.0	98.0	49.5	9.3	98.1	30.2	29.0	97.6	53.2
SMOTE	35.3	97.5	58.7	35.3	97.7	58.7	41.2	89.1	60.6
BLS	32.8	98.1	56.7	26.1	98.4	50.6	37.0	94.8	59.2
SLS	33.6	97.9	57.3	34.5	97.8	58.1	37.8	93.9	59.6
Cải tiến	53.1	91.2	**69.6**	59.2	81.1	**69.3**	60.0	75.3	**67.2**

Hình 3. Biểu đồ so sánh giá trị G-mean



Từ Bảng 2 và Biểu đồ tại Hình 3 nhận thấy rằng khi chúng ta áp dụng các phương pháp xử lý dữ liệu mất cân bằng thì hiệu quả phân lớp đều tốt hơn so với thực hiện trên dữ liệu gốc ban đầu, mặc dù độ đo *Specificity* có giảm nhẹ nhưng hai độ đo quan trọng là *Sensitivity* và *G-mean* đều tăng, điều này là rất có ý nghĩa với mô hình phân lớp dữ liệu mất cân bằng. Đặc biệt là phương pháp cải tiến của chúng tôi là có hiệu quả cao hơn với 3 phương pháp SMOTE, BLS, SLS trong cả 3 lần thực nghiệm với các thuật toán RF, SVM, KNN. Ngoài ra, theo Biểu đồ so sánh giá trị *G-mean* thì chúng ta nhận thấy rằng mặc dù phương pháp BLS và SLS là những biến thể của SMOTE nhưng hiệu quả phân lớp trong bài toán dự đoán khách hàng rời bỏ dịch vụ lại không cao hơn so với SMOTE.

4. Kết luận

Dự đoán khách hàng rời bỏ dịch vụ là một bài toán phân lớp dữ liệu mất cân bằng nên khi sử dụng các phương pháp xử lý dữ liệu mất cân bằng như SMOTE hay các biến thể của SMOTE nói chung là cho hiệu quả phân lớp tốt hơn so với dữ liệu gốc ban đầu. Trong bài báo này chúng tôi đã đề xuất một phương pháp mới giảm phân bố mất cân bằng của dữ liệu dựa trên sự kết hợp giữa đối nhãn của các phần tử đa số bị nhiễu và sinh thêm phần tử nhân tạothiểu số trong vùng lân cận của các phần tử thiểu số ban đầu. Kết quả thực nghiệm cho thấy rằng phương pháp đề xuất của chúng tôi có hiệu quả phân lớp tốt hơn so với các phương pháp xử lý dữ liệu mất cân bằng phổ biến. Bài toán dự đoán khách hàng rời bỏ dịch vụ vẫn còn rất nhiều thách thức cần giải quyết, trong thời gian tới nhóm tác giả sẽ tiếp tục nghiên cứu thêm về chiến lược sinh thêm phần tử nhân tạo tối ưu hơnnhằm nâng cao hiệu quả dự đoán.

Tài liệu tham khảo:

Aamer Hanif and Noor Azhar (2017), Resolving class imbalanced and feature selection in customer churn dataset, 2017 International Conference on Frontiers of Information Technology, pp. 82-86.

Annisa Aditsania, Adiwijaya and Aldo Lionel Saonard (2017), Handling Imbalanced Data in Churn Prediction using ADASYN and Backpropagation Algorithm, 2017 3rd International Conference on Science in Information Technology (ICSITech), pp. 533-536.

Duyen Do, Phuc Huynh, Phuong Vo and Tu Vu (2017), Customer Churn Prediction in an Internet Service Provider, 2017 IEEE International Conference on Big Data (BIGDATA), pp. 3928-3933.

Kien Vu (2018), Dự đoán khách hàng rời bỏ trong ngành viễn thông, truy cập ngày 27/10/2018 từ <https://kienvu2368.medium.com/>.

H. Ali, M. N. M. Salleh, R. Saedudin, K. Hussain and M. F. Mushtaq (2019), "Imbalance class problems in data mining: a review," Indonesian Journal of Electrical Engineering and Computer Science, vol. 14, no. 3, pp. 1560-1571.

Nguyễn Ngọc Tuân (2016), Áp dụng kỹ thuật khai phá dữ liệu dự báo thuê bao rời mạng di động, Luận văn thạc sĩ - Đại học Công nghệ, ĐHQGHN.

Nghiêm Thị Toàn, Nghiêm Thị Lịch, Bùi Dương Hương, Đặng Xuân Thọ, "Mask: phương pháp mới nâng cao hiệu quả phát hiện gian lận tài chính", Tạp chí Khoa học và Kỹ thuật - Học viện KTQS, số 184 (06-2017), pp 5-17.

Uma R. Salunkhe and Suresh N. Mali (2018), A Hybrid Approach for Class Imbalance Problem in Customer Churn Prediction: A Novel Extension to Under-sampling, I.J. Intelligent Systems and Applications,5, pp. 71-81.

Yanmin Sun, Andrew K.C. Wong and Mohamed S. Kamel (2009), "Classification of imbalanced data: a review," International Journal of Pattern Recognition and Artificial Intelligence, vol. 23, p. 687-719.