KSI - PHƯƠNG PHÁP KẾT HỢP PHÂN CỤM VỚI BỘ LỌC TÁI LẤY MẪU ĐỂ LOẠI BỎ NHIỄU TRONG DỮ LIỆU MẤT CÂN BẰNG

Bùi Dương Hưng[*], **Vũ Văn Thỏa**[+], **Đặng Xuân Thọ**[**]
[*] *Trường Đại học Công đoàn*
[+] *Học viện Công nghệ Bưu chính Viễn thông*
[**] *Trường Đại học Sư phạm Hà Nội*

Abstract: Dữ liệu phân lớp thường có phân bố số lượng không đồng đều giữa các nhãn lớp, vấn đề này được gọi là phân lớp dữ liệu mất cân bằng và xuất hiện ngày càng nhiều trong các ứng dụng thực tế. Kỹ thuật sinh thêm phần tử nhân tạo (SMOTE) là một trong những phương pháp tiền xử lý dữ liệu được biết đến nhiều nhất để giải quyết bài toán này. Tuy nhiên, theo các nghiên cứu gần đây, số lượng phần tử mất cân bằng không phải là một vấn đề chính mà hiệu quả phân lớp còn bị giảm do các yếu tố khác như sự phân bố dữ liệu với sự xuất hiện của các phần tử nhiễu và các phần tử ở biên. Hạn chế nội tại của SMOTE là sinh thêm nhiều phần tử nhiễu dạng này. Một số nghiên cứu đã chỉ ra bộ lọc nhiễu kết hợp với SMOTE sẽ nâng cao hiệu quả phân lớp (SMOTE-IPF). Ở bài báo này, chúng tôi đề xuất phương pháp kết hợp phân cụm với bộ lọc tái lấy mẫu nhằm giải quyết tốt hơn vấn đề này. Kết quả thực nghiệm trên các bộ dữ liệu tổng hợp và dữ liệu chuẩn quốc tế UCI với các mức độ mất cân bằng đã chỉ ra phương pháp đề xuất nâng cao hiệu quả của thuật toán SMOTE và SMOTE-IPF.

Keywords **:** SMOTE, IPF, Over-Sampling, dữ liệu mất cân bằng, phân lớp.

I. GIỚI THIỆU

Ngày nay, với sự xuất hiện ngày càng quan trọng của dữ liệu lớn, nghiên cứu về xử lý và khai phá dữ liệu lớn trở thành một chủ đề nóng, thách thức các phương pháp học máy truyền thống với mong muốn nhanh, hiệu quả, và chính xác. Hiện nay chưa có một phương pháp hiệu quả nào khai phá các loại dữ liệu thực tế. Đặc biệt, một khó khăn nữa mà chúng ta cũng thường phải đối mặt là dữ liệu mất cân bằng. Cụ thể như xác định những giao dịch thẻ tín dụng gian lận [1], kiểm tra các xâm nhập mạng trái phép [2], phát hiện vết dầu loang từ hình ảnh vệ tinh [3], các chuẩn đoán, dự đoán trong y sinh học [4].. Các phương pháp phân lớp dữ liệu chuẩn truyền thống thường gặp nhiều khó khăn do việc học bị lệch sang lớp đa số, dẫn đến độ chính xác thấp khi dự đoán lớp thiểu số.

Một số giải pháp cho vấn đề phân lớp dữ liệu mất cân bằng được đưa ra là dựa trên mức độ dữ liệu và mức độ thuật toán. Ở cấp độ thuật toán, các giải pháp cố gắng cải tiến các thuật toán phân lớp truyền thống để tăng cường việc học với các mẫu trong lớp thiểu số. Cụ thể như một số thuật toán học dựa trên chi phí với việc đặt thêm trọng số cho lớp thiểu số [5], điều chỉnh xác xuất dự đoán ở lá đối với phương pháp cây quyết định [6], bổ sung thêm hằng số phạt khác nhau cho mỗi lớp hoặc điều chỉnh ranh giới phân lớp cải tiến thuật toán máy vector hỗ trợ. Ở cấp độ dữ liệu, mục đích là để cân bằng sự phân bố các lớp bởi việc điều chỉnh mẫu vùng dữ liệu theo hai hướng gồm giảm kích thước mẫu lớp đa số hoặc tăng kích thước mẫu lớp thiểu số. Trong đó, có một số phương pháp phổ biến được áp dụng như Condensed Nearest Neighbor Rule (CNN) [7], Neighborhood Cleaning Rule (NCL) [8], Tomek links [9], SMOTE [10], Borderline-SMOTE [11], Safe-level-SMOTE [12]. Ngoài ra, một số nghiên cứu khác sử dụng các bộ lọc như lọc tập hợp EF [13], lọc phân vùng IPF [14] kết hợp với các phương pháp sinh thêm phần tử nhằm nâng cao hiệu quả phân lớp. Cụ thể như phương pháp SMOTE-IPF [15] được giới thiệu năm 2015 nhằm xử lý nhiễu trong các phân lớp mất cân bằng.

Mặc dù các phương pháp trên đã có những hiệu quả nhất định đối với phân lớp dữ liệu mất cân bằng có phần tử nhiễu. Tuy nhiên, các phương pháp này vẫn có những hạn chế nhất định như: SMOTE có một số hạn chế liên quan đến sinh thêm phần tử "mù". Bởi việc sinh thêm các phần tử nhân tạo (ở lớp thiểu số) chỉ làm một cách hình thức và do đó những phần tử ở mỗi lớp có thể bị gần sát nhau. Trong khi các đặc tính khác của dữ liệu bị bỏ qua như sự phân bố của các phần tử ở lớp đa số và thiểu số ở từng vùng khác nhau. Từ đó, tác giả đề xuất mở rộng mới (KSI) của SMOTE-IPF thông qua việc phân cụm, nhằm xác định các cụm dữ liệu có những phần tử lớp là thiểu số ở toàn cục nhưng lại là phần tử chiếm đa số trong cục bộ cụm. Dựa vào đó chúng tôi có cơ chế sinh thêm phần tử nhân tạo một cách phù hợp hơn, nâng cao hiệu quả phân lớp dữ liệu hơn. Trước khi đi vào giới thiệu chi tiết phương pháp KSI ở phần III, phần II sẽ trình bày

Tác giả liên lạc: Bùi Dương Hưng
Email: hungbd@dhcd.edu.vn
Đến tòa soạn: 30/04/2019, chỉnh sửa: 17/5/2019, chấp nhận đăng: 24/5/2019

về tiêu chí đánh giá. Một số kết quả đạt được và đánh giá sẽ được trình bày trong phần IV, và cuối cùng là phần kết luận.

II. TIÊU CHÍ ĐÁNH GIÁ

Nhằm đánh giá hiệu quả giữa các phương pháp phân lớp dữ liệu, đầu tiên, chúng ta xác định ma trận nhầm lẫn đối với phân lớp dữ liệu nhị phân, như được chỉ ra trong Bảng 1, *TP* là số lượng phần tử lớp positive được dự đoán đúng, *FN* là số lượng phần tử thực sự là positive nhưng bị dự đoán nhầm là negative, *FP* là số lượng phần tử thực sự là negative nhưng bị dự đoán nhầm là positive, *TN* là số lượng phần tử lớp negative được dự đoán đúng.

Bảng 1. Ma trận nhầm lẫn

Nhãn dự đoán	**Nhãn thực tế**	
	Lớp Positive	Lớp Negative
Lớp Positive	True Positive *(TP)*	False Positive *(FP)*
Lớp Negative	False Negative *(FN)*	True Negative *(TN)*

Một số độ đo được xác định dựa trên ma trận nhầm lẫn [16]–[18]:

$$TP_{rate} = \frac{TP}{TP + FN}$$
$$TN_{rate} = \frac{TN}{TN + FP}$$
$$FP_{rate} = \frac{FP}{TN + FP}$$
$$FN_{rate} = \frac{FN}{TP + FN}$$

Độ chính xác của các thuật toán phân lớp truyền thống được mô tả như sau:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Tuy nhiên, đối với dữ liệu mất cân bằng, số lượng phần tử lớp negative lớn hơn rất nhiều các phần tử lớp positive nên ảnh hưởng của *TP* là rất nhỏ, dễ dàng bị bỏ qua. Do đó, độ chính xác, *accuracy*, thường không được sử dụng khi đánh giá phân lớp dữ liệu mất cân bằng. Thay vào đó, các nghiên cứu thường sử dụng độ đo *G-mean* như một chỉ số đánh giá hiệu năng phân lớp của mô hình trên tập dữ liệu mất cân bằng.

$$G - mean = \sqrt{\frac{TP}{TP + FN} x \frac{TN}{TN + FP}}$$

G-mean là độ đo khả năng phân lớp tổng quát của cả lớp positive và negative của mô hình phân lớp [15], [16], [19], [20]. Trong bài báo này, phần thực nghiệm chúng tôi sử dụng *G-mean* để đánh giá hiệu quả của mô hình phân lớp dữ liệu.

Bên cạnh đó, trong nghiên cứ này chúng tôi sử dụng thêm độ đo *AUC (Area Under the ROC Curve)* – là diện tích bên dưới đường cong *ROC (Receiver Operating Characteristic curve),* một cách phổ biến để đánh giá chất lượng của các mô hình phân lớp với hai tiêu chí dựa trên ma trận nhầm lẫn là TP_{rate} và FP_{rate}. *AUC* dao động trong giá trị từ 0 đến 1 [21]. Một mô hình có dự đoán sai 100% có *AUC* là 0,0; và dự đoán chính xác 100% có *AUC* là 1.0.

III. PHƯƠNG PHÁP

A. *Phương pháp SMOTE*

Thuật toán SMOTE (Synthetic Minority Over-sampling Technique) được đề xuất năm 2002, nhằm giải quyết vấn đề mất cân bằng dữ liệu [10]. Đây là một trong những cách tiếp cận nổi tiếng nhất do sự đơn giản và hiệu quả của nó.

Cụ thể SMOTE sinh thêm phần tử nhân tạo bằng cách như sau: đầu tiên tìm hàng xóm gần nhất của mỗi phần tử của lớp thiểu số; sau đó chọn ngẫu nhiên một trong số những hàng xóm gần nhất; cuối cùng sinh thêm phần tử nhân tạo trên đoạn thẳng nối phần tử đang xét và láng giềng được lựa chọn bằng cách tính độ lệch giữa véc tơ thuộc tính của phần tử lớp thiểu số đang xét và láng giềng của nó.

B. *Phương pháp IPF*

Phương pháp lọc phân vùng lặp lại IPF (Iterative-Partitioning Filter) [14] loại bỏ các trường hợp nhiễu bằng cách lặp đi lặp lại cho đến khi đạt được một tiêu chí dừng. Quá trình lặp sẽ dừng nếu, đối với một số lặp lặp đi lặp lại, số lượng các phần tử nhiễu được xác định trong mỗi lần lặp lại này ít hơn 1% kích thước của tập dữ liệu huấn luyện ban đầu. Các bước cơ bản của mỗi lần lặp là:

(1) Chia tập dữ liệu huấn luyện D_T hiện tại thành các tập hợp con bằng nhau.

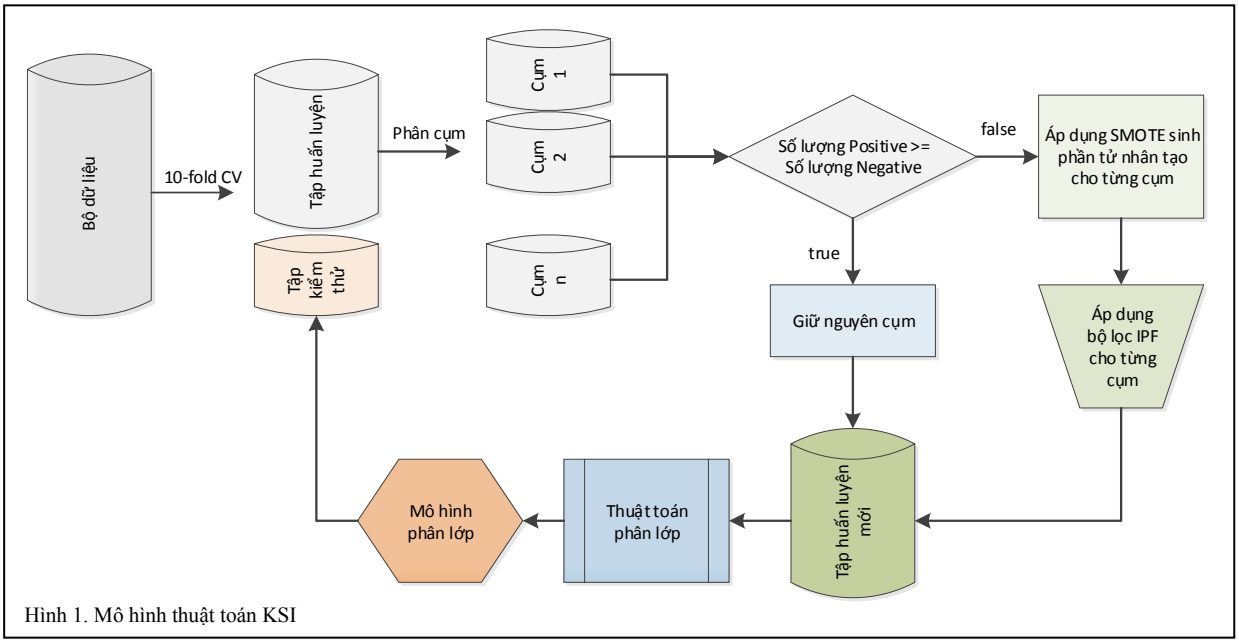
(2) Xây dựng mô hình với thuật toán C4.5 trên mỗi tập con này và sử dụng chúng để đánh giá toàn bộ tập dữ liệu huấn luyện hiện tại D_T.

(3) Thêm vào D_N các ví dụ nhiễu được xác định trong D_T sử dụng một chương trình bỏ phiếu.

(4) Loại bỏ nhiễu từ tập huấn luyện: $F_S = D_T \setminus D_N$

Quá trình lặp đi lặp lại kết thúc khi điều kiện dừng thỏa mãn, đó là, trong ba lần lặp lại liên tiếp, nếu số lượng các ví dụ nhiễu được xác định trong mỗi lần lặp là ít hơn 1% kích thước của các tập dữ liệu huấn luyện ban đầu, quá trình lặp đi lặp lại dừng.

C. *Phương pháp KSI*



Hình 1. Mô hình thuật toán KSI

Phương pháp SMOTE-IPF [15] được giới thiệu năm 2015 nhằm xử lý nhiễu trong các phân lớp mất cân bằng. Mặc dù SMOTE-IPF đã có những hiệu quả nhất định đối với mất cân bằng lớp có dữ liệu nhiễu, tuy nhiên phương pháp này vẫn có những hạn chế như: SMOTE có một số hạn chế liên quan đến sinh thêm phần tử "mù". Bởi việc sinh thêm các phần tử nhân tạo (ở lớp thiểu số) chỉ làm một cách hình thức và do đó những phần tử ở mỗi lớp có thể bị gần sát nhau. Trong khi các đặc tính khác của dữ liệu bị bỏ qua như sự phân bố của các phần tử ở lớp đa số và thiểu số ở từng vùng khác nhau, cụ thể như ở một số vùng dữ liệu, các phần tử lớp thiểu số ở toàn cục nhưng lại là phần tử chiếm đa số trong cục bộ vùng dữ liệu đó.

Từ đó, tác giả đề xuất mở rộng mới của SMOTE-IPF là thuật toán KSI (K-means-SMOTE-IPF) thông qua việc phân cụm, nhằm xác định các cụm dữ liệu có những phần tử lớp là thiểu số ở toàn cục nhưng lại là phần tử chiếm đa số trong cục bộ cụm. Dựa vào đó chúng tôi có cơ chế sinh thêm phần tử nhân tạo một cách phù hợp hơn, nâng cao hiệu quả phân lớp dữ liệu hơn. Mô hình thuật toán đề xuất KSI được mô tả chi tiết ở Hình 1. Đầu tiên, bộ dữ liệu được chia làm 10 phần, trong đó 9 phần làm tập huấn luyện, còn 1 phần làm tập kiểm thử. Sau đó, tập dữ liệu huấn luyện được phân cụm thành từng vùng dữ liệu nhằm kiểm tra mức độ mất cân bằng tại từng cụm cục bộ. Những cụm có phần tử lớp thiểu số ở toàn cục nhưng lại chiếm đa số tại cụm đó thì sẽ được giữ nguyên, không cần sinh thêm phần tử nhân tạo ở những vùng này. Ngược lại, ở những cụm các phần tử thiểu số ở toàn cục cũng là thiểu số ở cục bộ sẽ được áp dụng SMOTE và bộ lọc IPF. Cuối cùng chúng ta thu được tập dữ liệu mới. Chi tiết thuật toán KSI được mô tả như sau:

Input: Bộ dữ liệu huấn luyện (Train) gồm P phần tử thiểu số (positive) và N phần tử đa số (negative).

Output: Tập các phần tử nhân tạo thuộc lớp thiểu số.

Bảng 2. Bộ dữ liệu thực nghiệm

Dữ liệu	Số phần tử	Thuộc tính	Lớp thiểu số	Lớp đa số	Tỷ lệ mất cân bằng
abalone	731	8	42	689	1:16
blood	748	4	177	571	1:3
newthyroid	215	5	35	180	1:5
ecoli	768	8	268	500	1:8
haberman	306	3	81	225	1:3

Bước 1: Áp dụng thuật toán k-means để chia dữ liệu ban đầu (Train) thành các cụm *clust[1]*, *clust[2]*, *clust[3]*... *clust[n]*. Với N_i là tổng số phần tử đa số của cụm thứ i và P_i là tổng số phần tử lớp thiểu số của cụm thứ i trong đó *i = 1,2,3,...n*.

Bước 2: Trong tập dữ liệu (Train) có chứa các cụm *clust[i]* (với *i* là thứ tự các cụm *i = 1,2,3,...,n*) ta sẽ tiến hành lấy dữ liệu của *clust[1]*, *clust[2]*,..., *clust[n]*.

Bước 3: Xét điều kiện cần cho *clust[i]* để áp dụng thuật toán SMOTE. Ta gọi S_i là số phần tử nhân tạo sinh thêm trong cụm thứ i.

Nếu $N_i > P_i$ và $P_i > 5$ thì áp dụng thuật toán SMOTE cho clust[i] sinh ra S_i.

Nếu $N_i \leq P_i$ thì không áp dụng thuật toán SMOTE cho clust[i].

Nếu chứa nguyên N_i hoặc P_i thì không áp dụng thuật toán SMOTE cho clust[i].

Kết thúc bước 3, chúng ta thu được bộ dữ liệu $\{S_1, N_1, P_1, N_2, P_2, \dots, S_n, N_n, P_n\}$

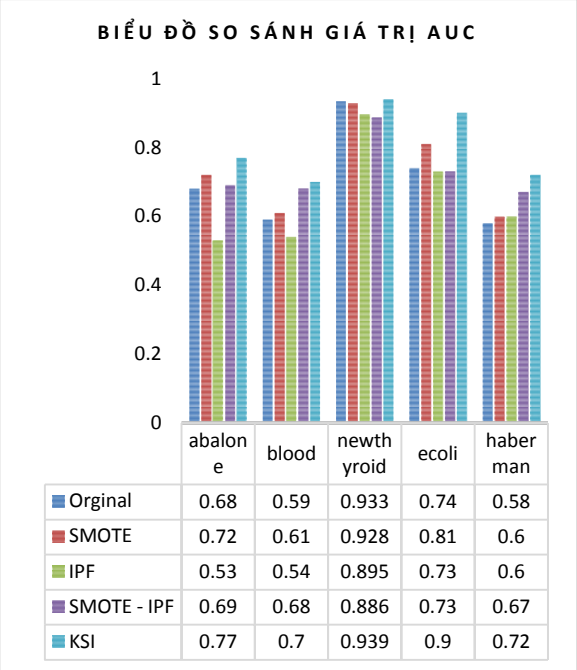
Bước 4: Sử dụng IPF để lọc dữ liệu dư thừa được sinh ra từ bước 3.

Bước 5: Dữ liệu sau khi được lọc bởi IPF được học để xây dựng mô hình. Kết thúc các bước của phương pháp đề xuất KSI.

IV. THỰC NGHIỆM VÀ ĐÁNH GIÁ

Các bộ dữ liệu được sử dụng là các bộ dữ liệu thực tế áp dụng cho phân lớp mất cân bằng với các phần tử nhiễu và đường biên, các bộ dữ liệu dành cho phân lớp mất cân bằng khác. Các bộ dữ liệu này có sẵn tại kho dữ liệu KEEL (http://keel.es) và kho dữ liệu UCI [22]. Cụ thể như sau ở Bảng 2.

Để đánh giá hiệu quả của phương pháp đề xuất KSI, chúng tôi đã tiến hành thực nghiệm trên các bộ dữ liệu được trình bày trong Bảng 2 với các phương pháp điều chỉnh dữ liệu: Original, SMOTE, IPF, SMOTE – IPF, và phương pháp KSI. Sau khi áp dụng các phương pháp điều chỉnh dữ liệu, các bộ dữ liệu mới được phân lớp bằng thuật toán phân lớp "bagging tree". Kết quả so sánh cuối cùng là giá trị trung bình của AUC và G-mean sau 20 lần thực hiện các phương pháp trên.

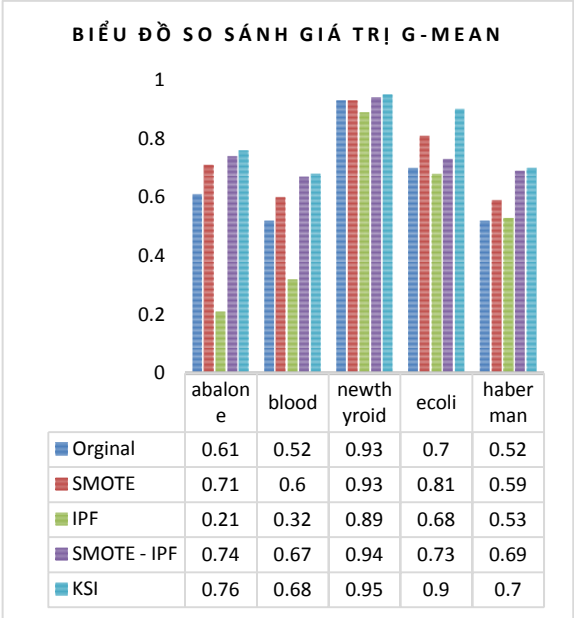


	abalone	blood	newthyroid	ecoli	haberman
Orginal	0.68	0.59	0.933	0.74	0.58
SMOTE	0.72	0.61	0.928	0.81	0.6
IPF	0.53	0.54	0.895	0.73	0.6
SMOTE - IPF	0.69	0.68	0.886	0.73	0.67
KSI	0.77	0.7	0.939	0.9	0.72

Hình 2. Biểu đồ so sánh giá trị AUC

Hình 2 và Hình 3 là các biểu đồ so sánh giá trị AUC và G-mean đánh giá kết quả thực hiện phân lớp trên mỗi bộ dữ liệu abalone, blood, newthyroid, ecoli và haberman khi chưa điều chỉnh (original) và khi đã được điều chỉnh bởi các thuật toán tiền xử lý SMOTE, IPF, SMOTE-IPF và KSI. Nhận thấy, với năm bộ dữ liệu, giá trị AUC của phương pháp đề xuất tốt hơn so với trường hợp dữ liệu ban đầu và các trường hợp dữ liệu áp dụng các thuật toán còn lại; với ba bộ dữ liệu blood, newthyroid, haberman, giá trị G-mean của phương pháp đề xuất tốt hơn; với hai bộ dữ liệu còn lại giá trị G-mean đạt kết quả cao hơn hẳn.

Cụ thể như với bộ dữ liệu abalone, độ đo AUC và G-mean của thuật toán KSI cũng được cải thiện hơn so với các thuật toán khác. Bộ dữ liệu abalone ban đầu có kết quả phân lớp AUC và G-mean chỉ đạt (68%, 61%). Các bộ dữ liệu sau khi được điều chỉnh đều có kết quả phân lớp được cải thiện đáng kể. Sau khi điều chỉnh bởi KSI, AUC cao nhất là 77%, G-mean đạt



	abalone	blood	newthyroid	ecoli	haberman
Orginal	0.61	0.52	0.93	0.7	0.52
SMOTE	0.71	0.6	0.93	0.81	0.59
IPF	0.21	0.32	0.89	0.68	0.53
SMOTE - IPF	0.74	0.67	0.94	0.73	0.69
KSI	0.76	0.68	0.95	0.9	0.7

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

76%. Tuy nhiên, nếu chỉ sử dụng bộ lọc IPF thì kết quả khá kém, AUC và G-mean chỉ đạt 53%, 21%. Điều này là do bộ lọc IPF trong quá trình lọc dữ liệu gốc đã loại bỏ đi một số dữ liệu gồm cả các phần tử lớp thiểu số, đây là những phần tử có ý nghĩa quan trọng trong phân lớp dữ liệu mất cân bằng.

Bên cạnh kết quả thực nghiệm với dữ liệu abalone, thuật toán đề xuất KSI cũng đạt hiệu quả rất tốt với bộ dữ liệu ecoli, cụ thể với độ đo AUC thuật toán KSI đã tăng hơn 16% so với dữ liệu ban đầu, và tăng hơn 9% so với thuật toán SMOTE. Với độ đo G-mean, phương pháp IPF không đạt hiệu quả mà còn làm giảm độ chính xác xuống 2%, tuy nhiên, thuật toán KSI đạt hiệu quả hơn hẳn dữ liệu ban đầu, SMOTE, IPF, và SMOTE-IPF lần lượt là (20%, 9%, 22%, và 17%).

V. KẾT LUẬN

Trong bài báo này, chúng tôi đã tập trung vào giải quyết của các phần tử nhiễu, đây là một vấn đề nghiên cứu quan trọng trong dữ liệu mất cân bằng. Đồng thời, chúng tôi nghiên cứu đề xuất thuật toán KSI mở rộng thuật toán SMOTE kết hợp với bộ lọc nhiễu IPF (SMOTE-IPF) nhằm kiểm soát tốt hơn các phần tử nhiễu được tạo ra bởi SMOTE. Sự phù hợp của cách tiếp cận trong phương pháp đề xuất đã được phân tích. Các kết quả thực nghiệm với độ đo AUC và G-mean đã chỉ ra rằng đề xuất KSI của chúng tôi có hiệu suất đáng chú ý hơn khi áp dụng vào các tập dữ liệu mất cân bằng với các phần tử nhiễu trên các bộ dữ liệu thực tế.

Mặc dù phương pháp KSI đã đạt được hiểu quả phân lớp tốt hơn so với một số phương pháp khác, vẫn còn nhiều chủ đề khác cần xem xét kỹ hơn trong hướng nghiên cứu này. Trong thời gian tới, chúng tôi nhận thấy có thể điều chỉnh cải tiến phương pháp KSI bằng cách áp dụng một số bộ lọc mới hiện nay như INFFC có thể cho kết quả lọc nhiễu tốt hơn bộ lọc

IPF, từ đó có thể nâng cao hiệu quả thuật toán phân lớp dữ liệu mất cân bằng. Bên cạnh đó, có thể kết hợp KSI với giảm chiều dữ liệu để áp dụng cho các bộ dữ liệu mất cân bằng có số lượng phần tử và thuộc tính lớn.

LỜI CẢM ƠN

Nghiên cứu này được hoàn thành dưới sự tài trợ của đề tài Nghiên cứu Khoa học cấp Bộ Giáo dục và Đào tạo Việt Nam, mã số đề tài B2018-SPH-52.

TÀI LIỆU THAM KHẢO

[1] M. Ahmed, A. N. Mahmood, and M. R. Islam, "A survey of anomaly detection techniques in financial domain," *Futur. Gener. Comput. Syst.*, vol. 55, no. January, pp. 278–288, 2016.

[2] M. Zareapoor, "Application of Credit Card Fraud Detection: Based on Bagging Ensemble Classifier," *Int. Conf. Intell. Comput. Commun. Converg.*, vol. 48, no. 12, pp. 679–686, 2015.

[3] G. Chen, Y. Li, G. Sun, and Y. Zhang, "Application of Deep Networks to Oil Spill Detection Using Polarimetric Synthetic Aperture Radar Images," *Appl. Sci.*, vol. 7, no. 10, p. 968, 2017.

[4] J. Jia, Z. Liu, X. Xiao, B. Liu, and K. C. Chou, "IPPBS-Opt: A sequence-based ensemble classifier for identifying protein-protein binding sites by optimizing imbalanced training datasets," *Molecules*, vol. 21, no. 1, 2016.

[5] Q. Cao and S. Wang, "Applying Over-sampling Technique Based on Data Density and Cost-sensitive SVM to Imbalanced Learning," 2011.

[6] F. Li, X. Zhang, X. Zhang, C. Du, Y. Xu, and Y.-C. Tian, "Cost-sensitive and hybrid-attribute measure multi-decision tree over imbalanced data sets," *Inf. Sci. (Ny).*, vol. 422, pp. 242–256, 2018.

[7] L. Si *et al.*, "FCNN-MR : A Parallel Instance Selection Method Based on Fast Condensed Nearest Neighbor Rule," *World Acad. Sci. Eng. Technol. Int. J. Inf. Commun. Eng.*, vol. 11, no. 7, pp. 855–861, 2017.

[8] M. Koziarski and M. Wozniak, "CCR: A combined cleaning and resampling algorithm for imbalanced data classification," *Int. J. Appl. Math. Comput. Sci.*, vol. 27, no. 4, pp. 727–736, 2017.

[9] M. Zeng, B. Zou, F. Wei, X. Liu, and L. Wang, "Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data," in *2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS)*, 2016, pp. 225–228.

[10] N. V Chawla, K. W. Bowyer, and L. O. Hall, "SMOTE : Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[11] H. Han, W. Wang, and B. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," *Lect. Notes Comput. Sci.*, vol. 3644, pp. 878–887, 2005.

[12] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique," *Lect. Notes Comput. Sci.*, vol. 5476, pp. 475–482, 2009.

[13] C. E. Brodley and M. A. Friedl, "Identifying mislabeled training data," *J. Artif. Intell. Res.*, vol. 11, pp. 131–167, 1999.

[14] T. M. Khoshgoftaar and P. Rebours, "Improving software quality prediction by noise filtering techniques," *J. Comput. Sci. Technol.*, vol. 22, no. 3, pp. 387–396, 2007.

[15] J. A. Sáez, J. Luengo, J. Stefanowski, and F. Herrera, "SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering," *Inf. Sci. (Ny).*, vol. 291, no. C, pp. 184–203, 2015.

[16] X. T. Dang, D. H. Tran, O. Hirose, and K. Satou, "SPY: A Novel Resampling Method for Improving Classification Performance in Imbalanced Data," in *2015 Seventh International Conference on Knowledge and Systems Engineering (KSE)*, 2015, pp. 280–285.

[17] A. Anand, G. Pugalenthi, G. B. Fogel, and P. N. Suganthan, "An approach for classification of highly imbalanced data using weighting and undersampling," *Amino Acids*, vol. 39, no. 5, pp. 1385–91, Nov. 2010.

[18] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets : A review," *Science (80-.).*, vol. 30, 2006.

[19] X. T. Dang *et al.*, "A novel over-sampling method and its application to miRNA prediction," *J. Biomed. Sci. Eng.*, vol. 06, no. 02, pp. 236–248, 2013.

[20] Z. Sun, Q. Song, X. Zhu, H. Sun, B. Xu, and Y. Zhou, "A novel ensemble method for classifying imbalanced data," *Pattern Recognit.*, vol. 48, no. 5, pp. 1623–1637, 2015.

[21] J. M. Lobo, A. Jiménez-valverde, and R. Real, "AUC: A misleading measure of the performance of predictive distribution models," *Glob. Ecol. Biogeogr.*, vol. 17, no. 2, pp. 145–151, 2008.

[22] E. K. T. Dheeru, Dua, "UCI Machine Learning Repository," *[http//archive.ics.uci.edu/ml]. Irvine, CA Univ. California, Sch. Inf. Comput. Sci.*, 2017.

KSI - A COMBINED CLUSTERING AND RESAMPLING METHOD WITH NOISE FILTERING ALGORITHM FOR IMBALANCED DATA CLASSIFICATION

Abstract: Classification datasets often have an unequal distribution of numbers between class labels, which is known as imbalance classification and appears more and more in real-world applications. SMOTE is one of the most well-known data-processing methods to solve this problem. However, as in recent researches, the imbalance distribution is not a main problem, the performance is reduced by other factors such as the distribution of data with the appearance of noisy samples. Some researchers have shown that SMOTE-based interference filters will improve efficiency (SMOTE-IPF). In this paper, we propose a clustering method with a re-sampling filter to archive better address this problem. Experimental results on UCI datasets with different levels of imbalance indicate the novel method improve the efficiency of the SMOTE and SMOTE-IPF algorithms.

Bùi Dương Hưng, Nhận học vị Thạc sỹ năm 2000. Hiện công tác tại Trường Đại học Công đoàn, nghiên cứu sinh khoá 2015, Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu: Khai phá dữ liệu, học máy.

Vũ Văn Thỏa, Nhận học vị Tiến sỹ năm 2002. Hiện công tác tại Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu: Công nghệ trí thức, điện toán đám mây, khai phá dữ liệu, xử lý ảnh, học máy.

Đặng Xuân Thọ, Nhận học vị Tiến sỹ năm 2013. Hiện công tác tại Khoa Công nghệ thông tin, Trường Đại học Sư phạm Hà Nội. Lĩnh vực nghiên cứu: Tin sinh học, khai phá dữ liệu, học máy.