Kỷ yếu Hội nghị Quốc gia lần thứ VIII về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Hà Nội, ngày 9-10/7/2015
DOI: 10.15625/vap.2015.000200

RANDOM BORDER UNDERSAMPLING: THUẬT TOÁN MỚI GIẢM PHẦN TỬ NGẪU NHIÊN TRÊN ĐƯỜNG BIÊN TRONG DỮ LIỆU MẤT CÂN BẰNG

Nguyễn Mai Phương[1], Trần Thị Ánh Tuyết[1], Nguyễn Thị Hồng[1], Đặng Xuân Thọ[1]
[1] Khoa Công nghệ thông tin, Trường Đại học Sư phạm Hà Nội
nguyenmaiphuong2710@gmail.com, tuyettran003@gmail.com, nguyenhong@hnue.edu.vn, thodx@hnue.edu.vn

TÓM TẮT— *Cùng với sự phát triển của lĩnh vực công nghệ thông tin là sự tăng lên nhanh chóng của dữ liệu. Dữ liệu càng lớn thì việc tìm ra những thông tin hữu ích trong đó càng trở nên khó khăn. Phân lớp dữ liệu là một trong những hướng nghiên cứu chính của khai phá dữ liệu. Phân lớp dữ liệu có ứng dụng nhiều trong thực tế, chẳng hạn như phát hiện thư rác, phát hiện xâm nhập mạng, phát hiện các gian lận giao dịch, chẩn đoán trong y học, phân tích hiệu quả điều trị. Tuy nhiên, nhiều bộ dữ liệu trong thực tế bị mất cân bằng tức là số phần tử giữa các lớp có sự chênh lệch. Việc phân lớp đúng các phần tử lớp thiểu số này lại đóng vai trò rất quan trọng. Ví dụ, trong cơ sở dữ liệu y học, số người mắc bệnh ung thư chiếm tỉ lệ rất nhỏ trên tổng số người bình thường. Việc chuẩn đoán sai người bị bệnh thành không bị bệnh có ảnh hưởng nghiêm trọng đến sức khỏe và tính mạng con người. Vì vậy cần tìm ra những phương pháp để cải thiện hiệu quả phân lớp dữ liệu. Bài báo này giới thiệu về mất cân bằng dữ liệu, sự cần thiết của việc phân lớp dữ liệu. Tiếp đó, chúng tôi xin đề xuất một thuật toán mới gọi là thuật toán giảm phần tử ngẫu nhiên trên đường biên trong dữ liệu mất cân bằng (Random Border Undersampling). Thuật toán được cải tiến từ thuật toán đã có Random undersampling, điểm mới là thuật toán không chỉ đơn giản là chọn các phần tử thuộc lớp đa số để giảm bớt mà xác định những phần tử trên đường biên của lớp đa số và giảm bớt các phần tử đó. Thuật toán được áp dụng vào các bộ dữ liệu từ nguồn dữ liệu chuẩn quốc tế UCI: Bộ dữ liệu Blood, Haberman, Pima và Breast-w, từ kết quả thực nghiệm cho thấy khả năng phân lớp dữ liệu được cải thiện.*

Từ khóa— *Dữ liệu mất cân bằng, Phân lớp dữ liệu, Random undersampling, Borderline, Random border undersampling*

I. GIỚI THIỆU

Hiện nay, bên cạnh sự phát triển mạnh mẽ của công nghệ thông tin và truyền thông là sự tăng lên của dữ liệu. Việc khai phá dữ liệu để chắt lọc những thông tin có ích càng trở nên khó khăn khi cơ sở dữ liệu ngày càng lớn. Phân lớp dữ liệu là một trong những hướng nghiên cứu chính của khai phá dữ liệu. Phân lớp là để dự đoán những nhãn lớp cho các bộ dữ liệu mới. Đầu vào là một tập các mẫu dữ liệu huấn luyện, với một nhãn phân lớp cho mỗi mẫu dữ liệu. Đầu ra là mô hình phân lớp dựa trên tập huấn luyện và các nhãn lớp.

Trong những năm qua, phân lớp dữ liệu đã thu hút sự quan tâm các nhà nghiên cứu trong nhiều lĩnh vực khác nhau. Công nghệ này cũng ứng dụng trong nhiều lĩnh vực khác nhau như: ngân hàng, y tế, dự báo thời tiết, thể thao, giải trí…Ví dụ, trong lĩnh vực ngân hàng, phân lớp dữ liệu dựa vào thông tin nghề nghiệp, thu nhập… của khách hàng sẽ đưa ra quyết định có cho khách hàng vay hay không. Trong dự báo thời tiết, có thể cho biết thời tiết ngày mai là mưa hay nắng dựa vào những thông số về độ ẩm, sức gió, nhiệt độ… của ngày hôm nay và các ngày trước đó. Trong chẩn đoán y khoa, phân lớp dữ liệu bệnh nhân để đưa ra chẩn đoán về khối u là lành tính hay ác tính.

Tuy nhiên, dữ liệu thu được trong thực tế thường mất cân bằng. Tập dữ liệu mất cân bằng thường xuất hiện trong các lĩnh vực như chẩn đoán y tế, giám sát hệ thống mạng, phát hiện xâm nhập hệ thống… Thông thường trong những lĩnh vực này lớp cần quan tâm lại có rất ít phần tử (minority) so với các lớp khác (majority) trong tập dữ liệu. Cụ thể như, trong số các trường hợp được xác định có bệnh hay không thì số người mắc bệnh là rất ít so với những người không bị bệnh [9]. Tuy nhiên việc xác định lớp thiểu số tức là số người mắc bệnh lại rất cần thiết và quan trọng. Việc mất cân bằng dữ liệu là một trong những lý do gây ra sự suy giảm về hiệu quả phân lớp của các thuật toán [8]. Dự đoán sai nhãn từ người "mắc bệnh" thành "không mắc bệnh" sẽ gây hậu quả nghiêm trọng đến tính mạng con người. Do vậy đây là một trong những vấn đề khó được cộng đồng và các nhà khoa học quan tâm nghiên cứu.

Tiếp theo, nội dung phần 2 chúng tôi xin trình bày về các nghiên cứu liên quan đến phương pháp tiếp cận trên mức độ dữ liệu trong dữ liệu mất cân bằng, phần 3 là giải thuật của thuật toán giảm phần tử ngẫu nhiên trên đường biên. Kết quả thực nghiệm sẽ được trình bày trong phần 4 và phần 5 là kết luận và hướng phát triển.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Trong các phương pháp tiếp cận trên mức độ dữ liệu, bằng cách điều chỉnh sự phân bố lớp để làm giảm sự mất cân bằng dữ liệu, thì thuật toán Under-sampling (giảm phần tử) và Over-sampling (tăng phần tử) là hai thuật toán phổ biến. Bên cạnh đó, người ta cũng có thể kết hợp cả 2 phương pháp trên, tức là cùng lúc giảm số phần tử ở lớp đa số và tăng phần tử lớp thiểu số [4].

A. Under-sampling (giảm phần tử)

Phương pháp giảm phần tử ở lớp đa số để làm giảm tính mất cân bằng dữ liệu. Cách đơn giản nhất đó là loại bỏ các phần tử ở lớp đa số một cách ngẫu nhiên. Ngoài ra cũng có một số cách giảm phần tử một cách có chủ đích như: giảm phần tử nhiễu ở vùng an toàn (safe level), giảm phần tử ở đường biên (borderline) [7].

Trong Random undersampling, ta ngẫu nhiên loại bỏ phần tử của lớp đa số trong tập dữ liệu huấn luyện cho đến khi ra tỉ số giữa đa số và lớp thiểu số phù hợp. Do đó, tổng số dữ liệu huấn luyện được giảm đáng kể.

B. Over-sampling (tăng phần tử)

Phương pháp tăng phần tử ở lớp thiểu số thì có 2 cách, cách thứ nhất là tăng phần tử lớp thiểu số bằng cách chọn ngẫu nhiên các phần tử lớp thiểu số sau đó sao chép giống hệt để làm tăng kích thước lớp thiểu số. Cách thứ 2 là tăng kích thước lớp thiểu số bằng cách sinh thêm các phần tử nhân tạo sau đó gán nhãn các phần tử đó thuộc lớp thiểu số.

III. THUẬT TOÁN MỚI RANDOM BORDER UNDERSAMPLING

A. Ý tưởng

Thuật toán Random Border Undersampling được cải tiến từ thuật toán Random undersampling đã có sẵn, sử dụng việc giảm ngẫu nhiên phần tử trên đường biên. Để xác định được các phần tử trên đường biên, thuật toán xác định dựa vào số láng giềng là thuộc lớp thiểu số m trong tổng số k láng giềng gần nhất. Nếu $k/2 \leq m < k$ thì phần tử đó là phần tử biên [1] [4].

B. Thuật toán Random Border Undersampling

Input: Bộ dữ liệu huấn luyện T gồm P positive lớp thiểu số, N negative lớp đa số

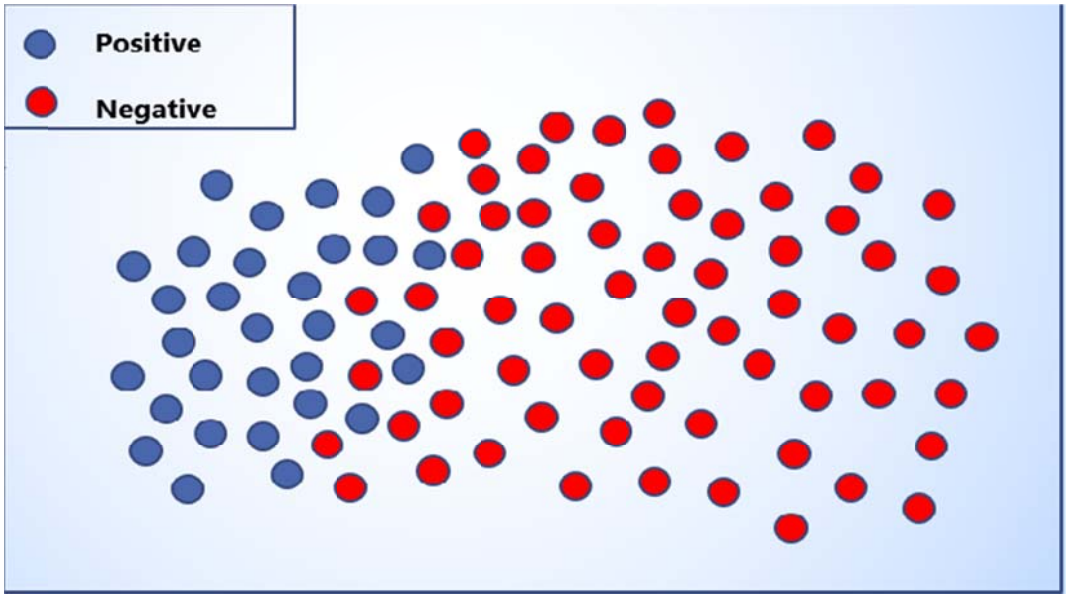
n: tỉ lệ phần trăm số phần tử trên biên bị giảm

k: số láng giềng gần nhất đối với một phần tử lớp đa số

m- số phần tử lớp đa số trên đường biên

Output: Tập dữ liệu đã giảm các phần tử trên biên theo tỉ lệ phần trăm n

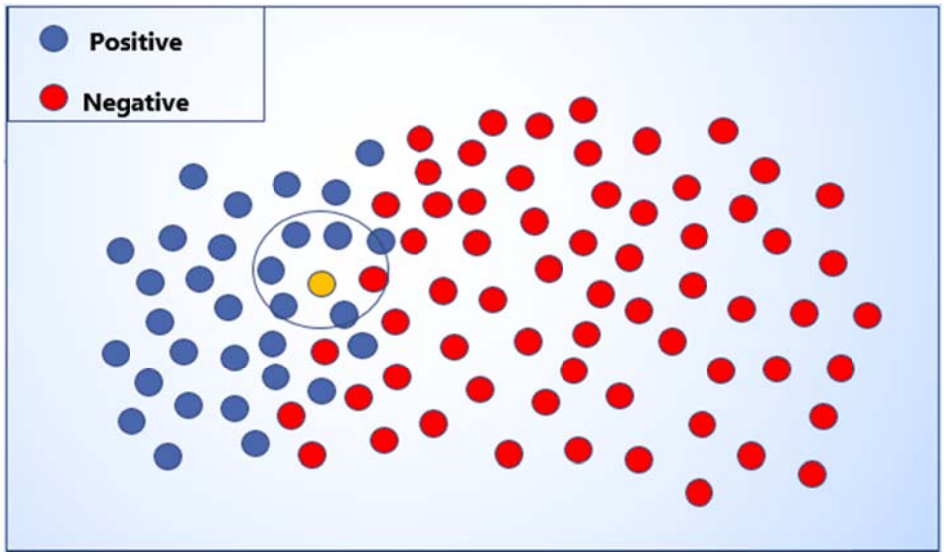
Dưới đây là những hình vẽ mô tả về bộ dữ liệu mất cân bằng và quá trình xác định phần tử lớp đa số thuộc đường biên để tiến hành giảm bớt các phần tử đó theo tỉ lệ phần trăm. Trong các hình vẽ dưới, hình tròn màu đỏ đại diện cho phần tử negative thuộc lớp đa số, hình tròn màu xanh da trời đại diện cho phần tử positive lớp thiểu số, hình tròn màu xanh lá cây là đại diện cho phần tử biên thuộc lớp đa số. Nhìn vào hình vẽ 1 ta nhận thấy ngay sự chênh lệch về số phần tử giữa hai lớp đa số và lớp thiểu số. Hình vẽ chỉ là mô tả cho một vùng của bộ dữ liệu lớn.



Hình 1. Phân bố dữ liệu

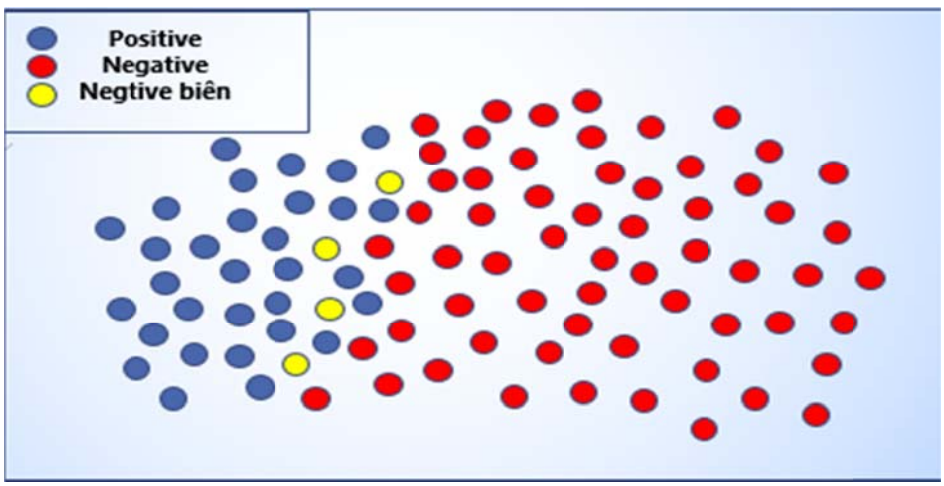
Hình 1 mô tả sự phân bố của dữ liệu, trong đó phần tử hình tròn màu đỏ là thuộc lớp đa số, được gán nhãn negative. Phần tử hình tròn màu xanh da trời là thuộc lớp thiểu số và được gán nhãn positive.

Tiếp đó, ở hình 2, ta đi xét k láng giềng cho từng phần tử lớp đa số. Giả sử trong k láng giềng gần nhất có m láng giềng là thuộc lớp thiểu số. Nếu m thỏa mãn điều kiện $\mathbf{k/2 \leq m < k}$ thì ta nói phần tử thuộc lớp đa số đang xét là phần tử thuộc đường biên. Ví dụ trên hình, phần tử đang xét là phần tử được đánh dấu màu vàng, trong số 7 láng giềng gần nhất của nó thì có 6 phần tử thuộc lớp thiểu số, 1 phần tử thuộc lớp đa số. Như vậy phần tử đó thuộc biên.

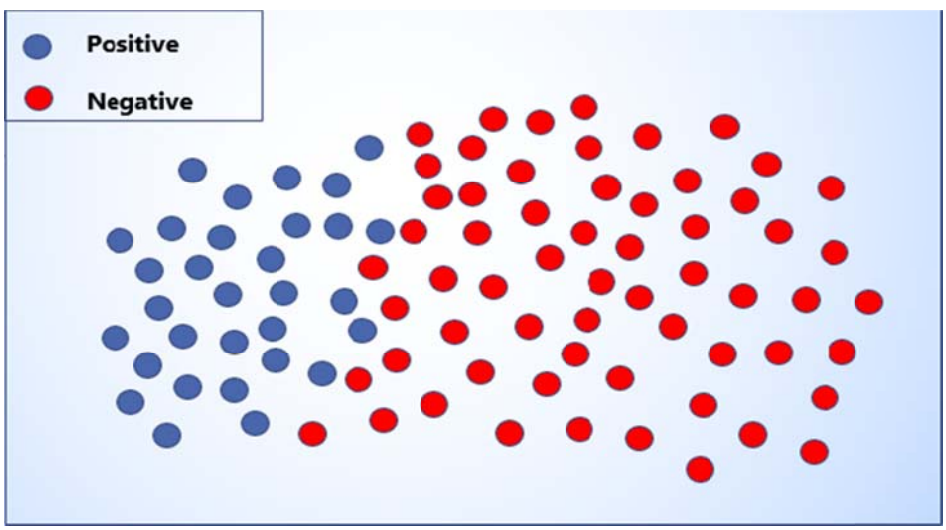


Hình 2. Xác định k-láng giềng

Tương tự , ta xác định được các phần tử lớp đa số thuộc đường biên là các phần tử màu vàng như trong hình 3.



Hình 3. Các phần tử biên



Hình 4. Xóa phần tử biên

Ta tiến hành xóa n% số phần tử biên thuộc lớp đa số đã xác định. Hình 4 là các phần tử thuộc biên đã bị xóa toàn bộ. Tuy nhiên, trong thuật toán mới của chúng tôi, số phần tử biên sẽ bị xóa theo tỉ lệ phần trăm, phụ thuộc vào tham số n.

Trong thuật toán, chúng tôi cho hai tham số đầu vào n chạy từ 1 đến 7 và k chạy từ 2 đến 10 để thuật toán khách quan và tổng quát hơn.

Các bước thực hiện của thuật toán:

Bước 1: Tìm các phần tử biên thuộc lớp đa số

Với mỗi phần tử trong tập lớp đa số N= $\{N_1, N_2, N_3, ... N_i\}$ ta tính k láng giềng gần nhất của nó trong toàn bộ tập dữ liệu huấn luyện T. Ta gọi số láng giềng thuộc lớp thiểu số trong tổng số k láng giềng gần nhất là m.

Bước 2: Xác định phần tử biên lớp đa số. Nếu **k/2 ≤ m < k** nghĩa là số láng giềng của N thuộc lớp thiểu số lớn hơn số láng giềng thuộc lớp đa số [7].

Bước 3: Ta đưa những phần tử thuộc N này vào một mảng border (mảng chứa các phần tử trên đường biên). Các phần tử trong mảng border là phần tử biên lớp đa số.

Bước 4: Ta giảm theo n phần trăm số phần tử trên đường biên để làm giảm mất cân bằng bộ dữ liệu.

Sau khi đã điều chỉnh bộ, ta áp dụng máy vecto hỗ trợ (Support Vector Machine) để phân lớp dữ liệu [9].

C. Áp dụng Máy véc tơ hỗ trợ (Support Vector Machine- SVM)

Support Vector Machines (SVMs) là một kỹ thuật học máy phổ biến, nhận dữ liệu vào và phân loại chúng vào hai lớp khác nhau cụ thể ở các bộ dữ liệu trong nghiên cứu của chúng tôi là lớp positive và negative. Do đó SVM là một thuật toán phân loại nhị phân. Tuy nhiên kỹ thuật SVM cũng được cho là hoạt động kém khi nó được áp dụng vào các bộ dữ liệu mất cân bằng [6]. Trong bài báo này, chúng tôi giải quyết vấn đề mất cân bằng dữ liệu dựa trên việc giảm phần tử trên đường biên. Chúng tôi sử dụng thuật toán SVM vì đây là thuật toán đem lại hiệu quả phân lớp tốt nhất trong các thuật toán hiện nay. Hơn nữa, thuật toán SVM là dựa vào khoảng cách lề cực đại (tức là đường phân tách chia hai lớp dữ liệu) [5], việc giảm các phần tử biên sẽ làm tăng khoảng cách phân chia giữa hai lớp nên sẽ đem lại hiệu quả phân lớp cao. Đồng thời, thực nghiệm của chúng tôi áp dụng SVMs cho cả RUS và phương pháp mới RBUS nên vẫn đảm bảo tính khách quan của kết quả thuật. Kết quả thực nghiệm cho thấy rằng giảm phần tử tốt hơn hẳn so với việc dùng bộ dữ liệu ban đầu.

IV. THỰC NGHIỆM

A. Bộ dữ liệu

Để đánh giá hiệu quả của thuật toán mới giảm phần tử ngẫu nhiên trên đường biên (RBUS), chúng tôi tiến hành thực nghiệm trên 4 bộ dữ liệu mất cân bằng lấy từ nguồn dữ liệu chuẩn quốc tế UCI. Kho dữ liệu UCI gồm các bộ dữ liệu của nhiều lĩnh vực. Trong bài báo này, chúng tôi tiến hành thực nghiệm trên các bộ dữ liệu thuộc lĩnh vực y sinh. Chi tiết các bộ dữ liệu như sau:

Bảng 1. Bốn bộ dữ liệu từ nguồn dữ liệu chuẩn quốc tế UCI

Bộ dữ liệu	Số phần tử	Số thuộc tính	Tỷ lệ mất cân bằng (Positive/ Negative)
Blood	748	4	1:3
Breast-w	699	8	1:2
Haberman	276	4	1:3
Pima	768	8	1:2

Các bộ dữ liệu trên đều có tình trạng mất cân bằng với hai nhãn lớp là Negative (lớp đa số) và Positive (lớp thiểu số). Tỉ lệ mất cân bằng của bộ Blood và Haberman là 1:3; bộ dữ liệu Breast-w và Pima là 1:2. Trong đó bộ dữ liệu có số phần tử lớn nhất là Pima với 768 phần tử và số thuộc tính là 8.

B. Các tiêu chí đánh giá

Để đánh giá hiệu quả của thuật toán, người ta căn cứ vào một số tiêu chí dựa trên ma trận nhầm lẫn

Bảng 2. Bảng ma trận nhầm lẫn

	Dự đoán là Possitive	Dự đoán là Negative
Thực tế là Positive	TP (số phần tử Positive dự đoán đúng)	FN (số phần tử Positive bị dự đoán sai)
Thực tế là Negative	FP (số phần tử Negative bị dự đoán sai)	TN (số phần tử Negative dự đoán đúng)

Một số tiêu chí đánh giá dựa trên ma trận nhầm lẫn **[3]:**

$$TP_{rate} = TP/(TP+FN)$$

$$TN_{rate} = TN/(TN+FP)$$

$$G\text{-}mean = \sqrt{TPrate.TNrate}$$

Gmean là một độ đo dùng để đánh giá hiệu quả của phân lớp dữ liệu mất cân bằng. Nếu tỉ lệ phân lớp đúng các phần tử ở 2 lớp cao thì G-mean sẽ cao.

C. Kết quả

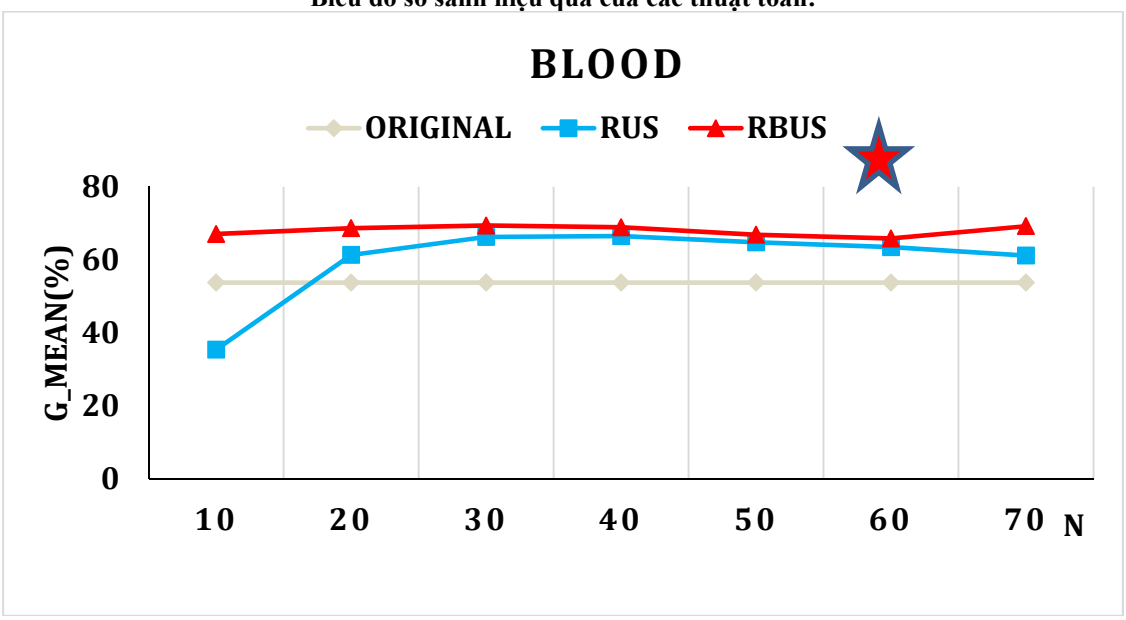
Chúng tôi so sánh kết quả phân lớp của dữ liệu sau khi áp dụng thuật toán mới Random Border Under-sampling với kết quả phân lớp của bộ dữ liệu ban đầu và bộ dữ liệu sau khi điều chỉnh bởi Random Undersampling. Thông tin các bộ dữ liệu được cho trên bảng 1. Với bộ dữ liệu ban đầu và các bộ dữ liệu sau khi điều chỉnh, chúng tôi thực hiện phân lớp bằng thuật toán SVM [6].

Bộ dữ liệu được chia làm 10 phần trong đó lần lượt mỗi phần được chọn làm bộ dữ liệu kiểm tra và 9 phần còn lại là bộ dữ liệu huấn luyện. Ta sẽ có 10 bộ đôi dữ liệu kiểm tra và huấn luyện tương ứng. Mỗi bộ dữ liệu huấn luyện được đưa vào thuật toán SVM thu được mô hình phân lớp. Sau khi có được mô hình phân lớp, bộ dữ liệu kiểm tra được đưa vào mô hình để xem có bao nhiêu mẫu được phân lớp đúng, bao nhiêu mẫu được phân lớp sai, từ đó ta tính được các độ đo. Các độ đo đánh giá của một lần 10-fold là trung bình cộng các độ đo đánh giá của 10 bộ dữ liệu huấn luyện và kiểm tra. Để kết quả thu được khách quan hơn, chúng tôi thực hiện 20 lần 10-fold vì việc chia dữ liệu là ngẫu nhiên. Hiệu quả phân lớp cuối cùng của mỗi bộ dữ liệu ứng với một thuật toán phân lớp được đánh giá thông qua các độ đo đánh giá là trung bình cộng của 20 lần 10-fold.

Để xác định thuật toán có ý nghĩa thống kê hay không, chúng tôi áp dụng kiểm định T-test. Nếu p-value của kiểm định này nhỏ hơn hoặc bằng 0.05 thì ta nói hai giá trị trung bình khác biệt và có ý nghĩa thống kê. Trong bài báo, chúng tôi sử dụng hàm t.test trong gói stats của R để tính giá trị

Dưới đây là các biểu đồ so sánh hiệu quả phân lớp của kỹ thuật phân lớp chuẩn đối với bộ dữ liệu ban đầu, bộ dữ liệu sau khi được điều chỉnh bằng thuật toán giảm phần tử ngẫu nhiên (RUS) và bộ dữ liệu sau khi được điều chỉnh bởi thuật toán mới giảm phần tử ngẫu nhiên trên đường biên (RBUS). Trục ngang là giá trị n- số phần trăm phần tử biên bị giảm. Trục dọc là giá trị Gmean tương ứng. Những biểu đồ có hình sao là thể hiện thuật toán có ý nghĩa thống kê.

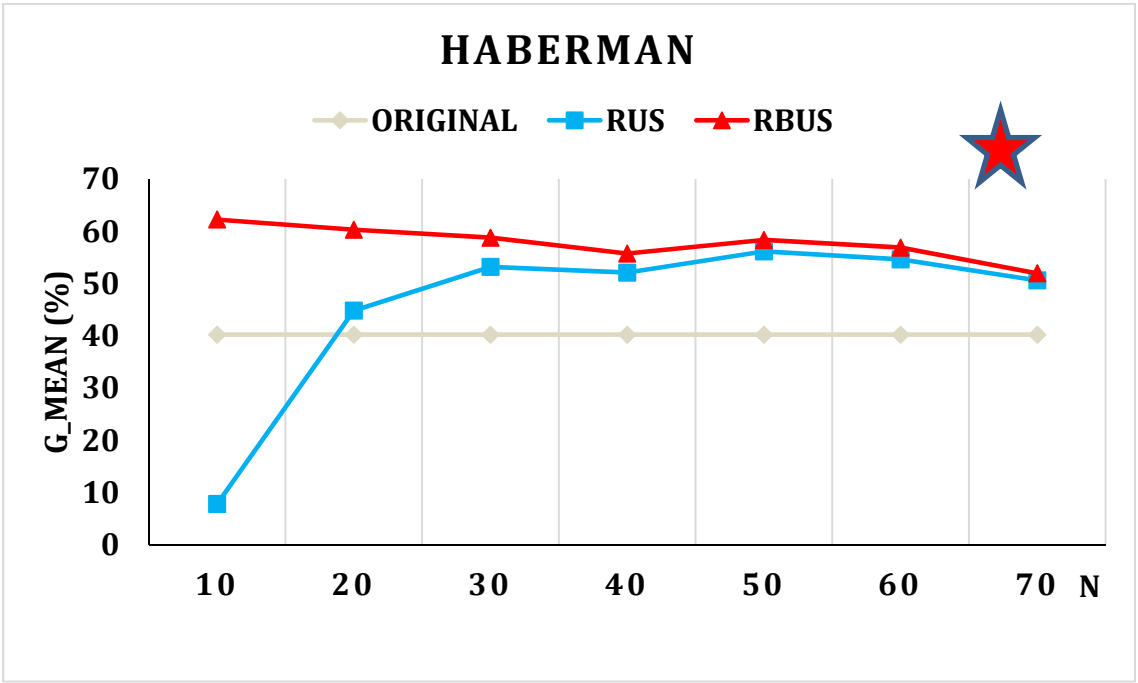
Biểu đồ so sánh hiệu quả của các thuật toán:



Hình 5. So sánh giá trị G-mean của bộ dữ liệu BLOOD

Trong bộ dữ liệu **BLOOD**: Giá trị G-mean cao nhất là tại n=70. So sánh giữa G-mean của bộ dữ liệu sau khi áp dụng thuật toán RBUS và với bộ dữ liệu ban đầu thì giá trị p-value đạt **2.2e-16,** với bộ dữ liệu áp dụng thuật toán RUS thì giá trị p-value đạt **2.693e-08.**

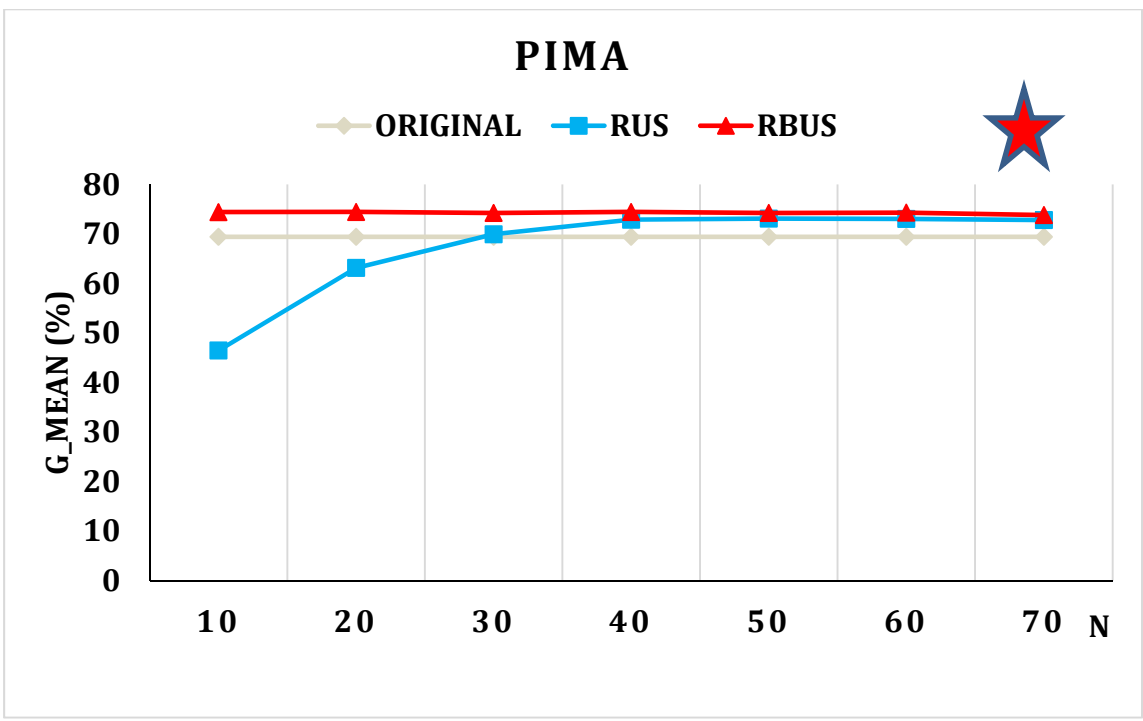
Bộ dữ liệu Blood khi áp dụng thuật toán mới RBUS cho kết quả P-value khi so sánh với bộ dữ liệu ban đầu và bộ dữ liệu được điều chỉnh bởi thuật toán RUS đều nhỏ hơn 0,05. Như vậy thuật toán mới đạt kết quả cao hơn và có ý nghĩa thống kê.



Hình 6. So sánh giá trị G-mean của bộ dữ liệu HABERMAN

Trong bộ dữ liệu **HABERMAN**: Giá trị G-mean cao nhất là tại n=10. So sánh giữa G-mean của bộ dữ liệu sau khi áp dụng thuật toán RBUS và với bộ dữ liệu ban đầu thì giá trị p-value đạt **2.2e-16,** với bộ dữ liệu áp dụng thuật toán RUS thì giá trị p-value đạt **8.021e-08.**

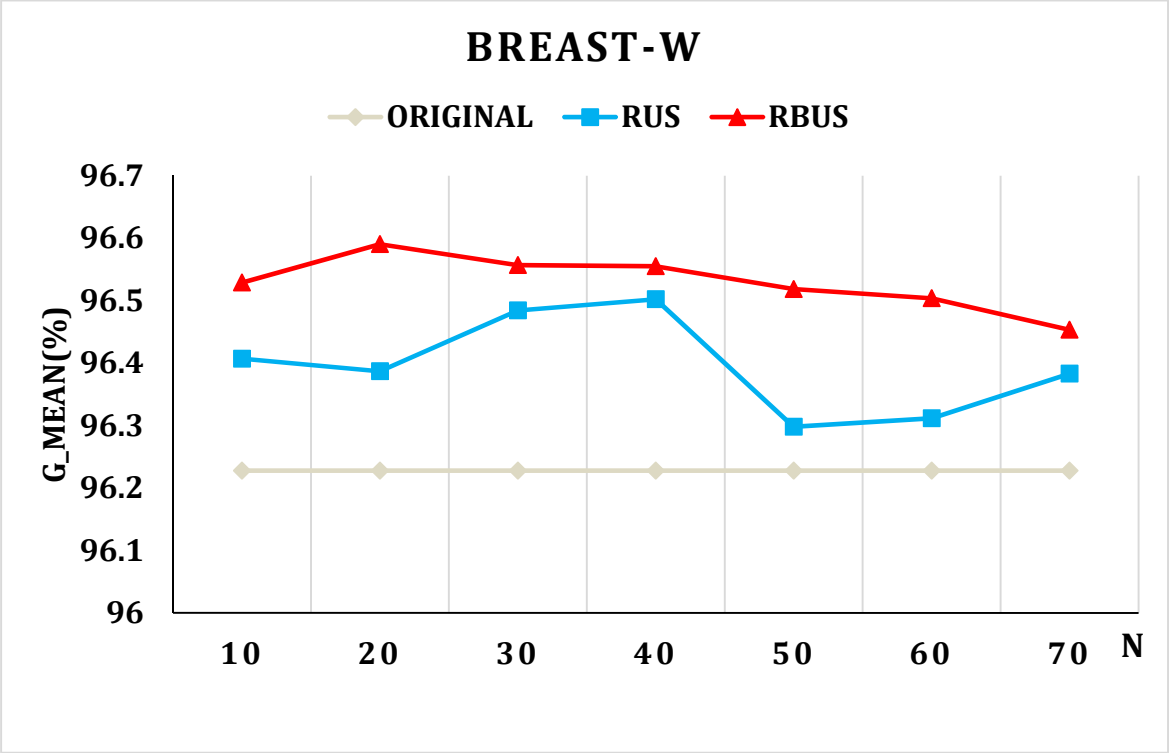
Bộ dữ liệu Haberman khi áp dụng thuật toán mới RBUS cho kết quả P-value khi so sánh với bộ dữ liệu ban đầu và bộ dữ liệu được điều chỉnh bởi thuật toán RUS đều nhỏ hơn 0,05. Như vậy thuật toán mới đạt kết quả cao hơn và có ý nghĩa thống kê.



Hình 7. So sánh giá trị G-mean của bộ dữ liệu PIMA

Trong bộ dữ liệu **PIMA**: Giá trị G-mean cao nhất là tại n=20. So sánh giữa G-mean của bộ dữ liệu sau khi áp dụng thuật toán RBUS và với bộ dữ liệu ban đầu thì giá trị p-value đạt **2.2e-16,** với bộ dữ liệu áp dụng thuật toán RUS thì giá trị p-value đạt **0.0019.**

Bộ dữ liệu Pima khi áp dụng thuật toán mới RBUS cho kết quả P-value khi so sánh với bộ dữ liệu ban đầu và bộ dữ liệu được điều chỉnh bởi thuật toán RUS đều nhỏ hơn 0,05. Như vậy thuật toán mới đạt kết quả cao hơn và có ý nghĩa thống kê.



Hình 8. So sánh giá trị G-mean của bộ dữ liệu BREAST-W

Trong bộ dữ liệu **BREAST-W**: Giá trị G-mean cao nhất là tại n=20. So sánh giữa G-mean của bộ dữ liệu sau khi áp dụng thuật toán RBUS và với bộ dữ liệu ban đầu thì giá trị p-value đạt **0.006124,** với bộ dữ liệu áp dụng thuật toán RUS thì giá trị p-value đạt **0.3194.**

Bộ dữ liệu Breast-w khi áp dụng thuật toán mới RBUS cho kết quả P-value khi so sánh với bộ dữ liệu ban đầu và bộ dữ liệu được điều chỉnh bởi thuật toán RUS lớn hơn 0,05. Như vậy thuật toán mới chưa có ý nghĩa thống kê. Mặc dù giá trị G-mean của thuật toán mới là lớn hơn so với thuật toán RUS và thuật toán SVM với bộ dữ liệu ban đầu nhưng nó chưa thực sự cách biệt nên cho giá trị p >0,05.

Bảng 3. Giá trị p-value

Bộ dữ liệu	Tên thuật toán	Random border undersampling
Blood	Original	< 2.2e-16
	Random undersampling	<2.693e-08
Breast-w	Original	0.006124
	Random undersampling	0.3194
Haberman	Original	<2.2e-16
	Random undersampling	<8.021e-08
Pima	Original	<2.2e-16
	Random undersampling	0.0019

Trên đây là bảng các giá trị p-value để so sánh giá trị G-mean đánh giá hiệu quả phân lớp trên các bộ dữ liệu Blood, Breast-w, Pima và Haberman sau khi được điều chỉnh bằng thuật toán Random Borderline Undersampling với các bộ dữ liệu ban đầu và các bộ dữ liệu sau khi được điều chỉnh bởi thuật toán Random Undersampling. Những giá trị p-value <0.05 (có ý nghĩa thống kê) nghĩa là hiệu quả phân lớp sau khi điều chỉnh dữ liệu bởi Random Borderline Undersampling tốt hơn hẳn.

Dưới đây là bảng thống kê số phần tử biên lớp đa số và tỉ lệ số phần tử biên lớp đa số so với tổng số phần tử lớp đa số của mỗi bộ dữ liệu:

Bảng 4. Thống kê số phần tử biên lớp đa số

Bộ dữ liệu	Số Negative (a)	Số Negative biên (b)	Negative biên : Negative (b:a)
Haberman	226	101	1:2
Blood	570	224	1:2
Pima	500	241	1:2
Breast-w	458	10	1:45

V. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Từ các kết quả thực nghiệm ở trên, chúng tôi thấy rằng sau khi điều chỉnh bộ dữ liệu bởi thuật toán mới Random Border Undersampling thì hiệu quả phân lớp các bộ dữ liệu cao hơn hẳn so với việc phân lớp của bộ dữ liệu ban đầu và bộ dữ liệu sau khi được điều chỉnh bởi Random Undersampling.

Từ bảng thống kê số phần tư biên lớp đa số và kết quả thực nghiệm, chúng tôi rút ra được kết luận rằng: khi áp dụng thuật toán mới Random Border Undersampling cho các bộ dữ liệu có nhiều số phần tử biên lớp đa số thì sẽ đạt hiệu quả phân lớp cao và đạt T-test. Trong 4 bộ dữ liệu chúng tôi thực nghiệm, có 3 bộ dữ liệu cho ra kết quả có ý nghĩa thống kê là Haberman, Blood và Pima. Ba bộ dữ liệu này đều có số phần tử biên lớp đa số chiếm xấp xỉ 50% trên tổng số các phần tử lớp đa số (b/a xấp xỉ 1/2). Bộ dữ liệu Breast-w có số phần tử biên lớp đa số là 10 trong tổng số 458 phần tử lớp đa số, tỉ lệ b/a là rất nhỏ và bằng 1:45 nên bộ Breast-w chỉ cho kết quả Gmean cao hơn so với Original và RUS nhưng chưa đạt ý nghĩa thống kê.

Trong thời gian tới, chúng tôi sẽ tiếp tục tìm hiểu và nghiên cứu vấn đề mất cân bằng dữ liệu. Chúng tôi cũng sẽ nghiên cứu thêm về thuật toán giảm phần tử nhiễu trên vùng an toàn (Safe level undersampling) và đặc biệt là kết hợp 2 thuật toán Random Border Oversampling và Random Border Undersampling để đưa ra hướng giải quyết tốt hơn cho việc xử lý dữ liệu mất cân bằng.

VI. TÀI LIỆU THAM KHẢO

[1]. Nguyễn Thị Hồng, Nguyễn Mạnh Cường, Đặng Xuân Thọ. Add-border-SMOTE: Phương pháp mới sinh thêm phần tử trong dữ liệu mất cân bằng*, Tạp chí Khoa học và Kỹ thuật - Học viện KTQS* - Số 164 (10-2014).

[2]. Nguyễn Thị Thùy Linh –Trường Đại học Quốc gia Hà Nội. Khóa luận tốt nghiệp: Nghiên cứu các thuật toán phân lớp dữ liệu dựa trên cây quyết định, *Hà Nội năm 2005*

[3]. Y. Sun, A. K. C. Wong, and M. S. Kamel, Classification of Imbalanced data: A review, *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 4, pp. 687–719, 2009.

[4]. Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline SMOTE A New Over-Sampling Method in Imbalanced Data Sets Learning, *J. Artificial Intell. Res*, 2002.

[5]. Man Sun Kim- An effective Under-sampling method for class imbalance data problem.

[6]. Rehan Akbani, Stephen Kwek, Nathalie Japkowicz. Applying Support Vector Machines to Imbalance Datasets, ECML 2004, pp. 39-50, 2014.

[7]. Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou, Senior Member, IEEE. Exploratory Undersampling for Class-Imbalance Learning, 6th IEEE International Conference onData Mining (ICDM'06), 2006, 965-969.

[8]. Bùi Minh Quân, Phạm Xuân Hiền, Huỳnh Xuân Diệp. Nâng cao độ chính xác phân loại lớp ít mẫu từ tập dữ liệu mất cân bằng*, Tạp chí Khoa học Trường đại học Cần Thơ,* 2013

[9]. Jiawei Han, Micheline Kamber and Jian Pei*, Data Mining: concepts and techniques, The Morgan Kaufmann Series of ELSEVIER,* 2012.