RANDOM BORDER-OVER-SAMPLING: THUẬT TOÁN MỚI SINH THÊM PHẦN TỬ NGẪU NHIÊN TRÊN ĐƯỜNG BIÊN TRONG DỮ LIỆU MẤT CÂN BẰNG

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

Tóm tắt: Phân lớp dữ liệu mất cân bằng là bài toán quan trọng xuất hiện trong hầu hết các lĩnh vực, đặc biệt là trong y sinh học chuẩn đoán người bệnh. Hiện nay, đã có nhiều nghiên cứu giải quyết bài toán này, trong đó, phương pháp tiền xử lý dữ liệu như Random Over-Sampling (ROS) là một phương pháp phổ biến và cho kết quả tốt. Tuy nhiên, một số trường hợp ROS lại không đạt được kết quả như mong đợi hoặc giảm hiệu quả phân lớp. Chính vì vậy, bài báo này tập trung nghiên cứu cải tiến thuật toán ROS, từ đó, đề xuất thuật toán mới Random Border-Over-Sampling (RBOS) bằng việc chọn các phần tử thiểu số có ý nghĩa quan trọng trên đường biên. Kết quả thực nghiệm trên sáu tập dữ liệu mất cân bằng từ nguồn dữ liệu chuẩn quốc tế UCI (*breast-p, blood, pima, haberman, glass, và coil2000*) đã chỉ ra thuật toán mới đề xuất của chúng tôi đạt hiệu quả tốt hơn hẳn so với phương pháp trước.

Từ khóa: Border-line, Random-Sampling, Over-Sampling, dữ liệu mất cân bằng, phân lớp.

I. MỞ ĐẦU

Ngày nay, trong thực tế xuất hiện rất nhiều bộ dữ liệu mất cân bằng, điển hình như: việc phát hiện tràn dầu trên bề mặt đại dương dựa vào các hình ảnh thu được từ rada vệ tinh, những hình ảnh có sự cố tràn dầu là rất nhỏ trong tổng số hình ảnh thu được, nên việc phát hiện chúng là rất khó, khiến cho công tác hạn chế ô nhiễm môi trường gặp nhiều khó khăn. Trong y học [1]–[3], số người mắc bệnh ung thư chiếm tỉ lệ rất nhỏ trên tổng số người dân, nhưng việc chuẩn đoán nhầm người bị bệnh thành người không bị bệnh có ảnh hưởng nghiêm trọng đến tính mạng con người. Trong giao dịch tín dụng hoặc cước di động, số giao dịch gian lận là rất nhỏ trên tổng số giao dịch, đặc biệt việc không phát hiện được hay phát hiện nhầm những giao dịch gian lận có thể gây thiệt hại lớn về tài chính đối với các doanh nghiệp [4]. Tại Hoa Kỳ, việc gian lận cước di động tiêu tốn hàng trăm triệu đô la mỗi năm.

Bài toán phân lớp dữ liệu đã được nghiên cứu với rất nhiều thuật toán phân lớp chuẩn như máy véc tơ hỗ trợ (SVM), k láng giềng gần nhất (K-NN), cây quyết định.. Tuy nhiên, khi xuất hiện các dữ liệu mất cân bằng, các thuật toán chuẩn trên không cho hiệu quả phân lớp cao như mong muốn. Chính vì vậy, yêu cầu đặt ra cần có phương pháp phân lớp phù hợp đối với các tập dữ liệu mất cân bằng nhằm đáp ứng các yêu cầu thực tế ngày càng tăng.

Nhiều công trình nghiên cứu trong và ngoài nước [5]–[9] đã giải quyết bài toán phân lớp dữ liệu mất cân bằng theo nhiều hướng khác nhau, theo các hướng tiếp cận ở cấp độ dữ liệu [10]–[13] và tiếp cận ở cấp độ thuật toán [14]–[17]. Trong đó, ở nghiên cứu này, chúng tôi tập trung vào hướng tiếp cận ở cấp độ dữ liệu, tiền xử lý dữ liệu để làm giảm sự mất cân bằng dữ liệu trước khi áp dụng các phương pháp phân lớp chuẩn nhằm mục đích cho hiệu quả tích cực. Điều chỉnh dữ liệu cũng có nhiều cách: giảm kích thước mẫu dữ liệu hoặc tăng kích thước mẫu dữ liệu. Thuật toán đại diện cho kỹ thuật này là Random Over-Sampling (ROS) và Random Under-Sampling (RUS). Ngoài ra, có thể kết hợp cả hai phương pháp trên để nâng cao hiệu quả phân lớp. Ramdom Over-Sampling là một phương pháp điều chỉnh tăng kích thước mẫu, thuật toán này sẽ lựa chọn ngẫu nhiên các phần tử trong lớp thiểu số và nhân bản chúng, làm cho bộ dữ liệu giảm bớt sự mất cân bằng. Ngoài ra, cũng có một số cách sinh phần tử có chủ đích như: tăng phần tử thiểu số ở vùng an toàn (Safe level), tăng phần tử ở đường biên (Borderline) [18]… Phương pháp điều chỉnh giảm kích thước mẫu Random Under-Sampling sẽ loại bỏ các phần tử ở lớp đa số một cách ngẫu nhiên đến khi tỷ số giữa các phần tử lớp thiểu số và các phần tử lớp đa số phù hợp. Do đó, số lượng các phần tử lớp đa số của tập huấn luyện sẽ giảm đáng kể.

Hai phương pháp trên được thực nghiệm chứng minh là hiệu quả, cải tạo tính mất cân bằng dữ liệu nhanh chóng. Tính ngẫu nhiên đảm bảo tính khách quan nhưng vẫn tồn tại một vài nhược điểm, trong một số trường hợp vẫn chưa đạt kết quả mong muốn. Phần tiếp theo của bài báo chúng tôi đề xuất nghiên cứu cải thiện thuật toán Random Over-Sampling thành thuật toán mới có tên Random Border-Over-Sampling nhằm

Tác giả liên hệ: Bùi Dương Hưng,
email: hungbd@dhcd.edu.vn
Đến tòa soạn: 06/2017, chỉnh sửa: 08/2017, chấp nhận: 09/2017

sinh các phần tử tập trung trên đường biên để nâng cao hiệu quả phân lớp, và được chứng minh bằng thực nghiệm trên các bộ dữ liệu chuẩn khác nhau.

II. GIẢI QUYẾT VẤN ĐỀ

A. *Mục tiêu nghiên cứu*

Qua tìm hiểu và nghiên cứu, chúng tôi nhận thấy ý nghĩa, tầm quan trọng của bài toán phân lớp dữ liệu mất cân bằng và những hạn chế mà thuật toán Random Over-Sampling (ROS) còn gặp phải là: Thứ nhất, việc nhân bản ngẫu nhiên làm tăng khả năng quá khít của mô hình phân lớp với bộ dữ liệu huấn luyện và làm tăng thời gian học nếu bộ dữ liệu huấn luyện ban đầu đã có kích thước lớn. Thứ hai, trong nhiều trường hợp có thể xảy ra tình trạng có những phần tử được chọn nhiều lần để tạo bản sao, cũng có những phần tử không được nhân bản lần nào. Nếu những phần tử không được lựa chọn để nhân bản lại là những phần tử có ích cho việc xây dựng mô hình phân lớp thì hiệu quả thuật toán cũng có thể bị giảm đi. Đặc biệt, trong một số nghiên cứu chỉ ra rằng các phần tử nằm trên đường biên giữa hai nhãn lớp dữ liệu đóng vai trò quan trọng trong quá trình phân lớp dữ liệu.

Chính vì vậy, chúng tôi đề xuất thuật toán mới Random Border-Over-Sampling (RBOS) với mục tiêu sinh thêm các phần tử nhân tạo trên đường biên nhằm khắc phục những hạn chế của thuật toán ROS hỗ trợ nâng cao hiệu quả phân lớp dữ liệu mất cân bằng.

B. *Thuật toán mới Random Border-Over-Sampling*

Trong bài toán phân lớp dữ liệu mất cân bằng, nhiều nghiên cứu đã chỉ ra rằng các thuật toán phân lớp và các thuật toán tiền xử lý dữ liệu cố gắng để xác định được đường phân chia ranh giới giữa hai lớp càng chính xác càng tốt. Đường phân chia ranh giới đó được gọi là đường biên của hai lớp. Phần tử biên (nằm trên hoặc gần đường biên) sẽ nằm gần với các phần tử lớp khác nhiều hơn so với những phần từ nằm xa biên. Vì thế, những phần tử này thường có khả năng bị gán nhãn hay bị phân lớp sai cao hơn so với những phần tử xa biên. Do đó, chúng có vai trò quan trọng trong việc quyết định hiệu quả phân lớp.

Trong bài báo khoa học [18], [19], nhóm tác giả Hui Han, Wen-Yuan Wang, and Bing-Huan Mao cũng đã khẳng định vai trò quan trọng của các phần tử biên thuộc lớp thiểu số trong việc phân lớp. Để xác định một phần tử lớp thiểu số có phải là phần tử nằm trên biên hay không, thuật toán xác định dựa vào số láng giềng thuộc lớp đa số m trong tổng số k láng giềng gần nhất. Nếu $k/2 \le m < k$ thì phần tử lớp thiểu số đó là phần tử biên và ngược lại [18], [19]. Bên cạnh đó, phương pháp ROS nguyên bản tác động bình đẳng trên tất cả phần tử lớp thiểu số, mà không quan tâm nhiều đến những phần tử trên đường biên nên khiến cho việc xác định những phần tử này gặp khó khăn. Chính vì vậy, chúng tôi đề xuất thuật toán mới RBOS, được mô tả như sau:

Thuật toán:

Input: Bộ dữ liệu huấn luyện T, trong đó, tập các phần tử lớp thiểu số D; $n\%$: tỉ lệ phần trăm số phần tử trên biên sinh thêm; k: số láng giềng gần nhất đối với một phần tử lớp thiểu số; m: số phần tử lớp đa số trong k láng giềng gần nhất bên trên.

Output: Bộ dữ liệu huấn luyện T và tập các phần tử sinh ngẫu nhiên trên đường biên D'

$D' = \emptyset$

$\forall p \in D$: tính k láng giềng gần nhất của p trong T

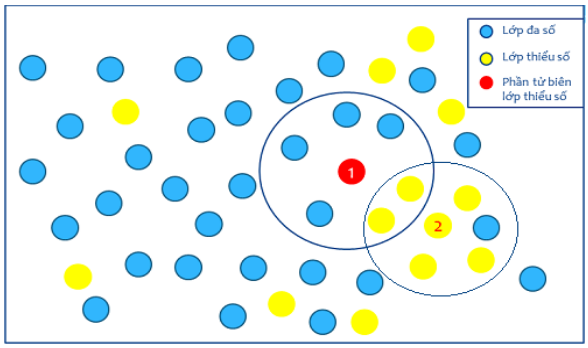
Tính số láng giềng thuộc lớp đa số trong số k láng giềng bên trên gọi là m

Nếu ($k/2 \le m < k$) thì p là phần tử biên của lớp thiểu số.

Thực hiện sinh thêm các phần tử trên đường biên theo tỉ lệ $n\% \in [100,200,..500]$.

return D'

Thuật toán RBOS khác so với ROS ở việc nhân bản có mục tiêu là những phần tử biên lớp thiểu số. Cách xác định một phần tử có là phần tử biên của lớp thiểu số hay không được minh họa bằng hình vẽ trực quan sau:



Hình 1. Cách xác định một phần tử biên lớp thiểu số

Trong hình 1, xét hai phần tử lớp thiểu số được đánh số 1 và 2, chọn ra sáu láng giềng gần nhất của chúng. Ta thấy, đối với phần tử số 1, trong sáu láng giềng gần nhất của nó có tới bốn phần tử thuộc lớp đa số và hai phần tử thuộc lớp thiểu số, khi đó, thỏa mãn điều kiện ($k/2 \le m < k$), vậy phần tử 1 là phần tử biên của lớp thiểu số và được lựa chọn để tạo ra phần tử nhân tạo. Tuy nhiên, đối với phần tử số 2, trong sáu láng giềng của nó chỉ có một phần tử lớp đa số, còn lại năm phần tử lớp thiểu số. Vì vậy, phần tử 2 không là phần tử biên và không được lựa chọn để tạo ra phần tử nhân tạo.

III. THỰC NGHIỆM

A. *Dữ liệu*

Để đánh giá hiệu quả thuật toán của thuật toán đề xuất, Random Border-Over-Sampling, chúng tôi tiến hành thực nghiệm trên sáu tập dữ liệu chuẩn quốc tế UCI [20]. Bảng I là thông tin về một số bộ dữ liệu mà bài báo nghiên cứu khoa học sử dụng trong quá trình thực nghiệm, đây đều là các bộ dữ liệu có sự mất cân bằng giữa các lớp.

Bảng I. Dữ liệu chuẩn quốc tế từ UCI

Dữ liệu	Số phần tử	Số thuộc tính	Tỉ lệ mất cân bằng
breast-p	198	32	1 : 4

blood	748	4	1 : 3
pima	768	8	1 : 2
haberman	306	3	1 : 3
glass	214	9	1 : 6
coil2000	5822	86	1 : 16

Dữ liệu được gán nhãn nhị phân gồm hai lớp, lớp lớp đa số được gán nhãn là Negative và thiểu số được gán nhãn là Positive. Trong đó, bộ dữ liệu coil2000 có tỉ lệ mất cân bằng lớn nhất là 1:16; bộ dữ liệu glass có tỉ lệ mất cân bằng là 1:6; bộ dữ liệu breast-p có tỉ lệ mất cân bằng là 1:4; bộ dữ liệu blood, haberman cùng có tỉ lệ mất cân bằng là 1:3; và bộ dữ liệu pima có tỉ lệ mất cân bằng là 1:2.

B. Các tiêu chí đánh giá

Để đánh giá một thuật toán phân lớp có hiệu quả hay không đều cần có những tiêu chí đánh giá cần thiết. Các tiêu chí đánh giá phân lớp được xây dựng trên cơ sở ma trận nhầm lẫn như minh họa ở bảng II như sau [21], [22].

Bảng II. Ma trận nhầm lẫn

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

Bảng II mô tả sự phân bố nhầm lẫn giữa hai lớp: Positive là nhãn lớp của các phần tử lớp thiểu số, Negative là nhãn lớp của các phần tử lớp đa số. TP là số phần tử có nhãn lớp thực tế là Positive và cũng được dự đoán là Positive; FP là số phần tử có nhãn lớp thực tế là Negative nhưng được dự đoán là Positive; TN là số phần tử có nhãn lớp thực tế là Negative và cũng được dự đoán là Negative; FN là số phần tử có nhãn lớp thực tế là Positive nhưng được dự đoán là Negative.

Dựa vào bảng II, chúng ta xác định được một số tiêu chí đánh giá sau [21].

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

$$TP_{rate} = \frac{TP}{TP + FN}$$

$$TN_{rate} = \frac{TN}{TN + FP}$$

$$G - mean = \sqrt{TP_{rate} * TN_{rate}}$$

Đối với dữ liệu cân bằng, tức số lượng phần tử Positive và Negative là gần như tương đương nhau, người ta thường căn cứ vào Accuracy (độ đo chính xác) để đánh giá hiệu quả phân lớp. Tuy nhiên, trong dữ liệu mất cân bằng, việc đánh giá hiệu quả phân lớp dựa vào độ chính xác không còn đáng tin cậy bởi vì số lượng Negative lớn, số phần tử Negative được dự đoán đúng cao tức TN cao, Positive rất nhỏ nên nhiều phần tử bị dự đoán sai tức TP nhỏ. Khi đó, mặc dù TP rất nhỏ nhưng độ chính xác (Accuracy) của mô hình phân lớp vẫn rất cao. Trong khi đó, thực tế vẫn có nhiều Positive bị dự đoán sai. Vì vậy, độ đo Accuracy không còn tin cậy trong việc đánh giá hiệu quả phân lớp của các tập dữ liệu mất cân bằng.

Trong nhiều bài báo khoa học cùng lĩnh vực [9], [18], [23]–[25], cũng như trong bài báo này, chúng tôi đánh giá hiệu quả thuật toán căn cứ vào giá trị G-mean. Trong đó, G-mean là độ đo phản ánh sự cân bằng giữa hiệu quả dự đoán các phần tử ở cả hai lớp, dựa trên độ đo TPrate và TNrate.

C. Kết quả thực nghiệm và đánh giá

Thuật toán ROS và thuật toán đề xuất RBOS đều là hai thuật toán tiền xử lý dữ liệu được xây dựng trên ngôn ngữ R và Perl [26]. Trong R, chúng tôi sử dụng package kernlab – để đánh giá hiệu quả phân lớp của hai phương pháp với thuật toán phân lớp chuẩn SVM.

Đầu tiên, chúng tôi chia ngẫu nhiên bộ dữ liệu ban đầu bằng phương pháp kiểm tra chéo (cross-validation) ra làm 10-fold có kích thước xấp xỉ nhau. Việc đánh giá thực hiện 10 lần, mỗi lần lấy một fold làm tập kiểm tra, 9 folds còn lại sử dụng làm tập huấn luyện. Với mỗi lần lặp, từ bộ dữ liệu huấn luyện ban đầu, ta thực hiện áp dụng một trong hai thuật toán ROS, và RBOS được bộ dữ liệu huấn luyện mới. Áp dụng thuật toán phân lớp SVM bộ dữ liệu huấn luyện mới này để thu được mô hình phân lớp. Sau đó, mô hình được đưa vào đánh giá với tập dữ liệu kiểm tra. Từ đó, qua 10 lần lặp, hiệu quả phân lớp được xác định là trung bình cộng của 10 giá trị độ đo tính được ở mỗi lần.

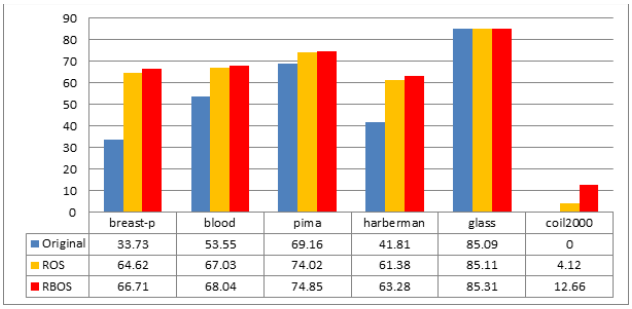
Cả hai thuật toán có tham số *n%* là số lần các phần tử lớp thiểu số ở vùng biên được chọn để nhân bản. Để tìm được kết quả tốt nhất chúng tôi cho chạy *n%* từ 100% đến 500%. Tương tự, với thuật toán mới RBOS, chúng tôi xét số láng giềng gần nhất *k* từ 3 đến 8 và lựa chọn kết quả tốt nhất.

Để kết quả được chính xác và khách quan, chúng tôi thực hiện 20 lần 10-fold, kết quả G-mean là giá trị trung bình của 20 lần thực hiện. Để kiểm tra xem G-mean của phương pháp nào thật sự cao hơn và có ý nghĩa thống kê, chúng tôi tiến hành kiểm định t-test. Kiểm định này sử dụng 20 lần chạy của G-mean cao nhất trong mỗi phương pháp. Kết quả của t-test là trị số xác suất p-value (probability value). Nếu p-value của kiểm định này nhỏ hơn hoặc bằng 0.05 thì hai giá trị trung bình khác biệt có ý nghĩa thống kê. Trường hợp ngược lại, p-value lớn hơn 0.05 thì hai giá trị trung bình khác biệt không có ý nghĩa thống kê [27]. Kiểm định này sử dụng hàm t.test trong gói stats của R để tính giá trị p-value.

Sau đây là kết quả G-mean thu được của sáu bộ dữ liệu khi thực hiện ba phương pháp phân lớp: phân lớp bằng thuật toán phân lớp chuẩn SVM trên bộ dữ liệu gốc (Original), phân lớp SVM kết hợp điều chỉnh mẫu bằng thuật toán ROS, và phân lớp SVM kết hợp điều chỉnh mẫu bằng thuật toán RBOS.

Có thể thấy, các kết quả thực nghiệm trên sáu bộ dữ liệu đã chỉ ra thuật toán đề xuất, RBOS, nâng cao hiệu quả phân lớp G-mean cao hơn so với phương pháp ROS và SVM trên dữ liệu nguyên gốc. Cụ thể, trong đó bộ dữ liệu breast-p đạt hiệu quả nổi bật khi G-mean đạt 66.71%, tăng 32.98% so với thuật toán SVM, và tăng 2.09% so với thuật toán ROS. Bộ dữ liệu

blood, khi áp dụng thuật toán mới RBOS thì giá trị G-mean thu được là 68.04% cao hơn so với phương pháp sử dụng thuật toán ROS có giá trị G-mean là 67.03%, và phương pháp chỉ chạy bộ dữ liệu gốc có giá trị G-mean là 53.55%. Và đặc biệt ở bộ dữ liệu coil2000, phương pháp ROS đã nâng cao hiệu quả phân lớp so với phương pháp chỉ chạy dữ liệu gốc có giá trị G-mean tăng từ 0% lên 4.12%, thì phương pháp mới đề xuất RBOS nâng cao hiệu quả vượt bậc khi đạt 12.66%.



Hình 2. Biểu đồ so sánh kết quả G-mean

Kết quả kiểm định t.test kiểm tra giá trị trung bình G-mean thu được khi áp dụng thuật toán RBOS so với thuật toán chuẩn và ROS của bộ dữ liệu breast-p, blood, pima, haberman và coil2000 cho giá trị p-value nhỏ hơn 0.05. Điều này cũng chỉ ra hiệu quả của thuật toán RBOS có ý nghĩa thống kê so với các thuật toán chỉ chạy bộ dữ liệu gốc và phương pháp ROS (chi tiết Bảng III).

Để làm rõ hơn vì sao chỉ có bộ dữ liệu Glass không đạt kiểm định t.test, chúng tôi tiến hành thống kê số lượng phần tử biên của lớp thiểu số (chi tiết bảng IV). Dựa trên kết quả bảng IV, chúng ta có thể dễ dàng nhận thấy trong sáu bộ dữ liệu, các bộ dữ liệu breast-p, blood, haberman, và coil2000 có tỉ lệ positive biên trên tổng số positive lần lượt là 91.30%, 77.40%, 76.25%, và 97.70%. Riêng bộ glass có tỉ lệ positive biên trên tổng số positive tương đối nhỏ, chỉ chiếm 17.9%, tức trong tổng số positive trên toàn tập dữ liệu, số positive biên là rất nhỏ. Đối chiếu với các kết quả đánh giá hiệu năng (G-mean) của các phương pháp ở hình 2, chúng ta có thể nhận thấy thuật toán đề xuất RBOS cho hiệu quả phân lớp tốt đối với các bộ dữ liệu có số positive biên lớn, cụ thể như breast-p, blood, haberman, và coil2000.

Bảng III. Thống kê kiểm định t-test

Dữ liệu	Thuật toán	ROS	RBOS	
breast-p		Original	< 2.2e-16	< 2.2e-16
		ROS		0.00222
blood		Original	< 2.2e-16	< 2.2e-16
		ROS		0.001453
pima		Original	< 2.2e-16	< 2.2e-16
		ROS		0.0005313
haberman		Original	< 2.2e-16	< 2.2e-16
		ROS		0.04733
glass		Original	0.4572	0.08127
		ROS		0.1806
coil2000		Original	1.18e-12	6.21e-15
		ROS		3.42e-05

Bảng IV. Thống kê số lượng phần tử biên của lớp đa số và lớp thiểu số

Dữ liệu	Số negative (a)	Số positive (b)	Số positive biên (c)	Tỉ lệ c/b (%)
breast-p	151	47	42	91.30
blood	570	178	137	77.40
pima	500	268	158	59.17
haberman	225	81	61	76.25
glass	185	29	5	17.85
coil2000	5474	348	340	97.70

IV. KẾT LUẬN

Trong bài báo này, chúng tôi đã trình bày tổng quan về bài toán phân lớp dữ liệu mất cân bằng là bài toán khó và nhiều thách thức nhưng có ý nghĩa lớn trong nghiên cứu và thực tế, từ đó, chúng tôi cũng đề xuất thuật toán mới dựa trên đường biên nhằm nâng cao hiệu quả phân lớp dữ liệu. Các kết quả thực nghiệm đánh giá trên sáu bộ dữ liệu mất cân bằng chuẩn UCI (*breast-p, blood, pima, haberman, glass, và coil2000*) đã chỉ ra rằng thuật toán đề xuất Random Border-Over-Sampling cho hiệu quả phân lớp tốt hơn thuật toán phân lớp chuẩn và thuật toán Random Over-Sampling. Điều này khẳng định tầm quan trọng của các phần tử biên trong tập dữ liệu có ảnh hưởng tới quá trình phân lớp.

Tuy nhiên, đặc thù trong mỗi bộ dữ liệu sẽ có phân bổ dữ liệu khác nhau, có dữ liệu thì số lượng phần tử trên đường biên nhiều và ngược lại. Qua thống kê số lượng phần tử biên và thực nghiệm đánh giá cũng chỉ ra rằng thuật toán cải tiến Random Border-Over-Sampling cho hiệu quả phân lớp tốt ở lớp dữ liệu có số lượng phần tử lớp thiểu số trên biên lớn.

Hiện nay, chưa có một phương pháp nào tối ưu hơn hẳn cho tất cả các bộ dữ liệu thực tế và trong ngành khai phá dữ liệu thì đều chấp nhận điều này. Trên cơ sở nghiên cứu và các kết quả đạt được, chúng tôi nhận thấy có nhiều vấn đề cần được tiếp tục nghiên cứu. Trong tương lai, chúng tôi sẽ tiếp tục nghiên cứu cải thiện thuật toán Random Border-Over-Sampling bằng cách kết hợp với các phương pháp khác như giảm số lượng phần tử biên thuộc lớp thiểu số hay loại bỏ các phần tử nhiễu để thuật toán đạt hiệu quả tốt hơn.

TÀI LIỆU THAM KHẢO

[1] W. K. Han, "Effective sample selection for classification of pre-miRNAs.," Genet. Mol. Res., vol. 10, no. 1, pp. 506–18, Jan. 2011.

[2] Y.-N. Zhang, D.-J. Yu, S.-S. Li, Y.-X. Fan, Y. Huang, and H.-B. Shen, "Predicting protein-ATP binding sites from primary sequence through fusing bi-profile sampling of multi-view features," BMC Bioinformatics, vol. 13, no. 1, p. 118, Jan. 2012.

[3] J. S. Chauhan, N. K. Mishra, and G. P. S. Raghava, "Identification of ATP binding residues of a protein from its primary sequence," BMC Bioinformatics, vol. 10, p. 434, Jan. 2009.

[4] W. Wang, "A Re-sampling Method for Class Imbalance Learning with Credit Data," pp. 393–397, 2011.

[5] H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, 2009.

[6] C.-Y. Yu, L.-C. Chou, and D. T.-H. Chang, "Predicting protein-protein interactions in unbalanced data using the primary

structure of proteins," BMC Bioinformatics, vol. 11, p. 167, Jan. 2010.

[7] X. T. Dang, O. Hirose, D. Hung Bui, T. Saethang, V. Anh Tran, L. Anh T. Nguyen, T. Kien T. Le, M. Kubo, Y. Yamada, and K. Satou, "A Novel Over-Sampling Method and its Application to Cancer Classification from Gene Expression Data," Chem-Bio Informatics J., vol. 13, pp. 19–29, 2013.

[8] L. Chen, Z. Cai, and L. Chen, "A Novel Differential Evolution-Clustering Hybrid Resampling Algorithm on Imbalanced Datasets," 2010 Third Int. Conf. Knowl. Discov. Data Min., pp. 81–85, Jan. 2010.

[9] C. Beyan and R. B. Fisher, "Classifying Imbalanced Data Sets using Similarity Based Hierarchical Decomposition," Pattern Recognit., vol. 48, no. 5, pp. 1653–1672, 2014.

[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] 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.

[12] 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.

[13] Barua, "MWMOTE—majority weighted minority oversampling technique for imbalaced data set learning," pp. 1–30, 2012.

[14] D. H. Tran, T. H. Pham, K. Satou, and T. B. Ho, "Prediction of microRNA Hairpins using One-Class Support Vector Machines," 2nd Int. Conf. Bioinforma. Biomed. Eng., pp. 33–36, May 2008.

[15] Y. Lin, Y. Lee, and G. Wahba, "Support Vector Machines for Classification in Nonstandard Situations," Mach. Learn., vol. 46, no. 1–3, pp. 191–202, 2000.

[16] S. Vluymans, I. Triguero, C. Cornelis, and Y. Saeys, "EPRENNID: An evolutionary prototype reduction based ensemble for nearest neighbor classification of imbalanced data," Neurocomputing, vol. 216, pp. 596–610, 2016.

[17] R. Alejo, R. M. Valdovinos, V. García, and J. H. Pacheco-Sanchez, "A hybrid method to face class overlap and class imbalance on neural networks and multi-class scenarios," Pattern Recognit. Lett., vol. 34, no. 4, pp. 380–388, 2013.

[18] H. M. Nguyen, E. W. Cooper, and K. Kamei, "Borderline Over-sampling for Imbalanced Data Classification," pp. 24–29, 2009.

[19] 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.

[20] A. Frank and A. Asuncion, "UCI Machine Learning Repository," [http//archive.ics.uci.edu/ml]. Irvine, CA Univ. California, Sch. Inf. Comput. Sci., 2010.

[21] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of Imbalanced Data: A Review," Int. J. Pattern Recognit., vol. 23, no. 4, pp. 687–719, 2009.

[22] L. Li, J. Xu, D. Yang, X. Tan, and H. Wang, "Computational approaches for microRNA studies: a review.," Mamm. Genome, vol. 21, no. 1–2, pp. 1–12, Feb. 2010.

[23] S. Oh, M. S. Lee, and B. Zhang, "Ensemble Learning with Active Example Selection for Imbalanced Biomedical Data Classification," vol. 8, no. 2, pp. 316–325, 2011.

[24] W. Klement, S. Wilk, W. Michalowski, and S. Matwin, "Classifying Severely Imbalanced Data," pp. 258–264, 2011.

[25] J. Tian, H. Gu, and W. Liu, "Imbalanced classification using support vector machine ensemble," Neural Comput. Appl., vol. 20, no. 2, pp. 203–209, Mar. 2010.

[26] A. Karatzoglou and A. Smola, "kernlab – An S4 Package for Kernel Methods in R," J. Stat. Softw., vol. 11, no. 9, 2004.

[27] J. Winter, "Using the Student ' s t -test with extremely small sample sizes," Pr. Assessment, Res. Evalutaion, vol. 18, no. 10, pp. 1–12, 2013.

RANDOM BORDER-OVERSAMPLING: A NOVEL METHOD IN IMBALANCED DATA SETS LEARNING

Abstract: Classification of imbalance data is an important problem that arises in most areas, especially in biomedical diagnoses. Currently, there are many researches try to solve this problem, in which, preprocessing method such as Random Over-Sampling (ROS) is a popular method and gives high performance. However, in some cases, ROS does not achieve the expected results or reduces the efficiency of the classification. Thus, this paper focuses on the improvement of the ROS algorithm, and thereby proposing a new Random Border-Over-Sampling (RBOS) algorithm by selecting significant minority samples on the borderline. Experimental results on six imbalanced data sets from UCI international data source (*breast-p, blood, pima, haberman, glass, and coil2000*) have shown that our proposed algorithm is effective and better than the previous method.

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 1990. Hiện công tác tại: Khoa Quốc tế và Đào tạo sau Đại học, Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu: Lý thuyết thuật toán, tối ưu hoá, hệ thông tin địa lý, mạng viễn thông.

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.