PHƯƠNG PHÁP MỚI DỰA TRÊN VÙNG AN TOÀN NÂNG CAO HIỆU QUẢ PHÂN LỚP DỮ LIỆU MẤT CÂN BẰNG

BÙI DƯƠNG HƯNG [1]
NGUYỄN THỊ HỒNG [2] *, ĐẶNG XUÂN THỌ* [2]
[1]*Khoa Tin học, Trường Đại học Công đoàn*
[2]*Khoa Công nghệ Thông tin, Trường Đại học Sư phạm Hà Nội*
Email: thodx@hnue.edu.vn

Tóm tắt: Nghiên cứu bài toán phân lớp trong thực tế như chuẩn đoán y học, phát hiện sự cố tràn dầu, phát hiện gian lận kinh tế tài chính… ngày càng được nhiều nhà khoa học quan tâm vì tầm ảnh hưởng lớn của những lĩnh vực này tới con người. Tuy nhiên, nhiều nghiên cứu đã chỉ ra hiệu quả phân lớp của các bài toán này chưa cao do có sự chênh lệch về số lượng phần tử giữa các lớp dữ liệu. Một yêu cầu đặt ra là cần có những hướng tiếp cận mới đối với trường hợp dữ liệu mất cân bằng để tăng tính hiệu quả phân lớp chính xác của thuật toán phân lớp. Bài báo của chúng tôi đề xuất ba thuật toán mới dựa trên mức an toàn nhằm nâng cao hiệu quả phân lớp dữ liệu. Hai thuật toán, Random Safe Oversampling (RSO) và Random Safe Undersampling (RSU) cải tiến trực tiếp từ Random Oversamling và Random Undersampling. Thuật toán thứ ba, Random Safe Over-Undersampling (RSOU) là sự kết hợp của RSO và RSU nhằm đồng thời vừa tăng thêm các phần tử ở lớp thiểu số vừa loại bỏ các phần tử ở lớp đa số ở các vùng an toàn.

Từ khóa: Dữ liệu mất cân bằng; Phân lớp dữ liệu; Safe level; Random Oversampling; Random Undersampling; Random Safe Oversampling; Random Safe Undersampling

1. MỞ ĐẦU

Ngày nay, một số lượng lớn của dữ liệu được thu thập và lưu trữ trong các cơ sở dữ liệu ở khắp mọi nơi trên thế giới. Không khó để tìm được các cơ sở dữ liệu lên tới Terabytes trong các doanh nghiệp và các trung tâm nghiên cứu [1], [2]. Rất nhiều thông tin và kiến thức vô giá tiềm ẩn trong các cơ sở dữ liệu như vậy, mà chưa có phương pháp tự động hiệu quả để phân tách thông tin [3]. Trong suốt nhiều năm, nhiều thuật toán được tạo ra để phân tách những gì được gọi là "quặng vàng" của tri thức từ các tập dữ liệu lớn. Đặc biệt, trong đó vấn đề phân lớp mất cân bằng ngày càng phổ biến trong một số lượng lớn các lĩnh vực có tầm quan trọng đối với cộng đồng khai phá dữ liệu. Đây là một trong 10 vấn đề khó đang được cộng đồng học máy và khai phá dữ liệu quan tâm. Một số phương pháp khác nhau để tiếp cận vấn đề này như phân lớp dữ liệu; kết hợp quy tắc; phân cụm dữ liệu… [4], [5] Trong đó giải thuật Random Oversampling là một trong những phương pháp nổi tiếng và tổng quát để giải quyết vấn đề mất cân bằng. Bài báo này sẽ phân tích chi tiết về phương pháp nghiên cứu mới – thuật toán sinh thêm phần tử dựa vào cấp độ an toàn trong dữ liệu mất cân bằng. Thuật toán mới sinh ra dựa trên sự phát

Tạp chí Khoa học và Giáo dục, Trường Đại học Sư phạm Huế
ISSN 1859-1612, Số 01(45)/2018: tr. 149-158
Ngày nhận bài: 27/3/2017; Hoàn thành phản biện: 23/5/2017; Ngày nhận đăng: 14/6/2017

triển từ thuật toán Random Oversampling. Giải thuật mới được đưa ra nhằm làm tăng hiệu quả phân lớp so với thuật toán Random Oversampling.

Phân lớp là một nhiệm vụ quan trọng của mô hình kiểu mẫu. Một loạt các thuật toán học máy chẳng hạn như Cây quyết định (Decision tree); Mạng lưới thần kinh lan truyền ngược; Mạng Bayes; k-láng giềng gần nhất; Máy vector hỗ trợ (Support Vector Machine)… đã được phát triển tốt và áp dụng thành công trên nhiều lĩnh vực [6]. Tuy nhiên, sự mất cân bằng của một tập dữ liệu đã gặp phải khó khăn tương đối nghiêm trọng cho hầu hết các thuật toán học phân lớp. Khó khăn quan trọng của vấn đề phân lớp mất cân bằng và sự xuất hiện thường xuyên của nó trong các ứng dụng thực tế của học máy và khai thác dữ liệu đã thu hút được sự quan tâm nghiên cứu. Một số ví dụ minh hoạ cho vấn đề khai phá dữ liệu mất cân bằng như phát hiện gian lận thẻ tín dụng; chuẩn đoán y học, phát hiện sự xâm nhập mạng, phát hiện sự cố tràn dầu từ các hình ảnh radar trên bề mặt trái đất, công nghiệp viễn thông… [6]. Nhiều nghiên cứu đã chỉ ra rằng, với các tập dữ liệu mất cân bằng như vậy sẽ làm cho mô hình học phân lớp gặp nhiều khó khăn trong dự báo dữ liệu của lớp thiểu số. Chính vì vậy, cần phải có những hướng tiếp cận mới đối với việc khai phá dữ liệu dạng này.

Một tập dữ liệu được coi là mất cân bằng nếu một trong các lớp có số lượng các phần tử quá nhỏ để so sánh với các lớp khác. Trong bài báo này, chúng tôi chỉ đề cập tới trường hợp phân lớp nhị phân, nghĩa là dữ liệu chỉ có hai nhãn lớp, một lớp có số lượng phần tử nhỏ hơn được gọi là lớp thiểu số, và lớp có số phần tử lớn hơn được gọi là lớp đa số. Ví dụ, tập dữ liệu Mammography chứa 10.923 mẫu được gán nhãn "negative" (Không ung thư) và 260 mẫu được gán nhãn "positive" (Ung thư). Nhiều nghiên cứu đã chỉ ra rằng, với dữ liệu Mammography, các phần tử lớp đa số được phân lớp với độ chính xác gần 100% nhưng lớp thiểu số có độ chính xác chỉ 0-10%. Giả sử, một phân lớp đạt độ chính xác là 10% đối với lớp thiểu số, nghĩa là sẽ có 234 mẫu lớp thiểu số bị phân loại sai thành lớp đa số. Điều đó sẽ dẫn đến 234 mẫu bị ung thư nhưng được chuẩn đoán nhầm là không bị ung thư [7]. Như vậy, việc phân lớp nhầm sẽ gây hậu quả nghiêm trọng. Từ đó cho thấy vai trò của việc giải quyết bài toán mất cân dữ liệu và đây cũng là vấn đề quan trọng được nhiều nhà nghiên cứu trong lĩnh vực học máy, khai phá dữ liệu quan tâm.

2. NỘI DUNG NGHIÊN CỨU

Các phương pháp để giải quyết vấn đề mất cân bằng lớp có thể được chia thành 2 loại: Phương pháp tiếp cận trên mức độ dữ liệu và phương pháp tiếp cận dựa trên mức độ thuật toán. Ở cấp độ dữ liệu, mục đích là để cân bằng sự phân bố các lớp, bởi việc điều chỉnh mẫu vùng dữ liệu. Ở cấp độ thuật toán, các giải pháp cố gắng thích ứng sự tồn tại các thuật toán phân lớp để tăng cường việc học ở lớp thiểu số. Các thuật toán học dựa trên chi phí (Cost-sensitive learning) kết hợp tiếp cận chặt chẽ cả cấp độ thuật toán và dữ liệu. Một vài thuật toán boosting cũng được báo cáo như những kỹ thuật meta có thể ứng dụng tới hầu hết các thuật toán học phân lớp. Ý tưởng chung của một phương pháp boosting là giới thiệu các loại chi phí tới dataframe của việc học tới AdaBoost.

Phương pháp tiếp cận trên mức độ dữ liệu với mục đích cân bằng sự phân bố các lớp, bằng việc điều chỉnh mẫu vùng dữ liệu. Thuật toán tiêu biểu của phương pháp này là Random Oversampling - tăng ngẫu nhiên phần tử ở lớp thiểu số và Random Undersampling - giảm ngẫu nhiên phần tử ở lớp đa số. Ngoài ra, có thể kết hợp cả hai phương pháp trên để đạt được hiệu quả phân lớp mong muốn. Random Oversampling (RO) 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. Ngược lại, phương pháp Random Undersampling (RU) sẽ loại bỏ các phần tử ở lớp đa số một cách ngẫu nhiên đến khi tỷ lệ giữa các phần tử thiểu số và đa số đạt một mức độ nhất định. Do đó số lượng các phần tử của tập huấn luyện sẽ giảm đáng kể. Hai phương pháp trên đã được thực nghiệm chứng minh là tương đối tốt, nhưng trong một số trường hợp lại đạt kết quả chưa mong muốn. Vì vậy chúng tôi đã nghiên cứu cách thức tăng (giảm) phần tử của lớp mất cân bằng dựa trên một mức độ "an toàn". Từ đó, đề xuất một phương pháp mới vừa tăng số lượng các phần tử an toàn ở lớp thiểu số, vừa giảm các phần tử an toàn ở mức đa số.

2.1. Random Safe Oversampling (RSO)

2.1.1. Ý tưởng

Phát triển từ thuật toán RO cùng với khái niệm vùng an toàn [8], chúng tôi đề xuất thuật toán RSO là phương pháp sinh thêm phần tử an toàn ở lớp thiểu số một cách ngẫu nhiên. Nếu thuật toán RO lựa chọn ngẫu nhiên các phần tử trong phân lớp thiểu số để nhân bản, thì thuật toán mới sẽ tập trung lựa chọn những phần tử "an toàn" trong phân lớp thiểu số để nhân bản. Thuật toán tính toán cấp độ an toàn của mỗi đối tượng dựa trên số láng giềng gần nhất của các đối tượng thiểu số trước khi sinh thêm các phần tử mới. Bằng cách sinh thêm nhiều hơn các phần tử nhân tạo lớp thiểu số xung quanh cấp độ an toàn lớn hơn, các kết quả thực nghiệm đã chỉ ra phương pháp mới RSO đạt hiệu suất chính xác hơn so với trước và so với thuật toán RO gốc.

Trong giải thuật RSO, cấp độ an toàn *safe level positive (slp)* được định nghĩa như trong công thức số (1) [8]. Nếu cấp độ an toàn *safe level positive* của một đối tượng gần tới *0*, đối tượng đó gần với phần tử nhiễu, ngược lại nếu nó gần tới *k*, đối tượng đó nằm trong vùng an toàn. Mức độ an toàn của một phần tử positive được định nghĩa trong công thức số (2). Nó thường được chọn vị trí an toàn tới các phần tử sinh nhân tạo.

safe level positive (slp) = số láng giềng là lớp thiểu số trong k láng giềng gần nhất (1)

safe level area (slp_area) = slp của đối tượng thuộc lớp thiểu số / k láng giềng gần nhất của phần tử đang xét (2)

Giả sử *p* là một phần tử dữ liệu lớp thiểu số đang xét, thì *slp_area* là mức độ an toàn của phần tử đó. Đối tượng lớp thiểu số có được nhân bản hay không sẽ phụ thuộc vào tỉ lệ *slp_area*. Nếu *slp_area>0.5*, nghĩa là xung quanh phần tử thiểu số đang xét có nhiều phần tử cùng nhãn với nó, thì phần tử thiểu số đang xét được coi là an toàn. Ngược lại, nếu *slp_area<0.5*, nghĩa là xung quanh phần tử thiểu số đang xét không có nhiều phần

tử cũng nhãn với nó, hoặc có nhiều phần tử nhiễu, lúc này ta sẽ loại phần tử đang xét đó mà không nhân bản chúng lên.

2.1.2. Thuật toán RSO

Input: Bộ dữ liệu huấn luyện *T* trong đó có tập các phần tử lớp thiểu số *D.*

N%: Số % positive được nhân bản bởi thuật toán RSO.

k: Số láng giềng gần nhất của phần tử positive.

Output: Bộ dữ liệu huấn luyện mới *T'* gồm tập các phần tử nhân bản mới sinh thêm *D'*.

Các bước thực hiện của thuật toán như sau:

$D' = \emptyset$

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

slp = số lượng các positive trong k láng giềng gần nhất của p trong D

slp_area = slp/k

if (0.5 < slp_area ≤ 1)

Nhân bản N% phần tử p an toàn đang xét;

return D'

2.2. Random Safe Undersampling (RSU)

2.2.1. Ý tưởng

Kết hợp ý tưởng từ thuật toán RU và khái niệm vùng an toàn, thuật toán RSU sẽ ngẫu nhiên loại bỏ các phần tử an toàn ở lớp đa số. Tương tự thuật toán RSO, với thuật toán RSU chúng tôi cũng định nghĩa cấp độ an toàn *safe level negative (sln)* ở công thức (3) và mức độ an toàn của một phần tử negative được định nghĩa trong công thức (4) như sau:

safe level negative (sln) = số láng giềng là lớp đa số trong k láng giềng gần nhất (3)

safe level area (sln_area) = sln của đối tượng thuộc lớp đa số / k láng giềng gần nhất của phần tử đang xét (4)

Nếu tỷ lệ *sln_area* của một đối tượng *n* đang xét nằm trong khoảng từ *0.5 ÷ 1*, tức là nằm trong vùng an toàn, thì ta sẽ loại bỏ phần tử negative này ra khỏi bộ dữ liệu.

2.2.2. Thuật toán RSU

Input: Bộ dữ liệu huấn luyện *T* trong đó có tập các phần tử lớp đa số *C*.

M%: Số % negative bị loại bỏ bởi thuật toán RSU.

k: Số láng giềng gần nhất của phần tử negative.

Output: Bộ dữ liệu huấn luyện *T'* đã loại bỏ tập các phần tử negative an toàn *C'*.

Các bước thực hiện thuật toán:

$C' = \emptyset$

$\forall n \in C$: *Tính k láng giềng gần nhất của n*

sln = số lượng negative trong k láng giềng gần nhất của n trong C

slp_area = sln/k

if (0.5 < sln_area ≤ 1)

Loại bỏ M% phần tử n an toàn đang xét;

return C'

2.3. Random Safe Over and Undersampling (RSOU)

Kết hợp hai thuật toán RSO và RSU ở trên, chúng tôi đề xuất thuật toán mới RSOU sẽ dựa trên cấp độ an toàn và mức độ an toàn của các đối tượng để vừa sinh thêm các phần tử an toàn của lớp thiểu số, vừa loại bỏ các phần tử an toàn của lớp đa số.

Thuật toán cụ thể như sau:

Input: Bộ dữ liệu huấn luyện *T* trong đó

Tập các phần tử lớp thiểu số *D*, và tập các phần tử lớp đa số *C*.

N%: Số % positive được nhân bản bởi thuật toán RSO.

M%: Số % negative bị loại bỏ bởi thuật toán RSU

k: Số láng giềng gần nhất của phần tử positive.

Output: Bộ dữ liệu huấn luyện mới *T'* gồm tập các phần tử nhân bản mới sinh thêm *D'*.

$D' = \emptyset$; $C' = \emptyset$

If (class = "Positive")

Thực hiện thuật toán RSO

If (class = "Negative")

Thực hiện thuật toán RSU

Return $D' \cup C'$

2.4. Thực nghiệm

2.4.1. Các tiêu chí đánh giá

Phân lớp được đánh giá bởi ma trận nhầm lẫn như minh họa trong Bảng 1. Các dòng của bảng là nhãn lớp thực tế của một đối tượng, và các cột của bàng là nhãn lớp dự đoán của một đối tượng. Như vậy, nhãn lớp của phân lớp thiểu số gọi là postive và nhãn lớp của phân lớp đa số gọi là negative. TP là số phần tử có nhãn lớp thực tế là positive và cũng được mô hình phân lớp dự đoán là positive. FP là số phần tử có nhãn lớp thực tế là

negative nhưng được mô hình phân lớp dự đoán là positive. FN là số phần tử có nhãn lớp thực tế là positive nhưng được mô hình phân lớp dự đoán là negative. TN là số phần tử có nhãn lớp thực tế là negative và cũng được mô hình phân lớp dự đoán là negative.

Bảng 1. *Ma trận nhầm lẫn*

	Positive dự đoán	**Negative dự đoán**
Positive thực tế	TP	FN
Negative thực tế	FP	TN

Một số độ đo được định nghĩa dựa trên ma trận nhầm lẫn **Error! Reference source not found.**:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (5)$$

$$TP_{rate} = \frac{TP}{TP+FN} \quad (6)$$

$$TN_{rate} = \frac{TN}{TN+FP} \quad (7)$$

$$G-mean = \sqrt{TP_{rate} \cdot TN_{rate}} \quad (8)$$

Trong đó, *G-mean* là độ đo biểu diễn hiệu quả phân lớp của cả hai lớp thiểu số và lớp đa số [9]. *G-mean* được xác định dựa vào hai giá trị TP_{rate} và TN_{rate}. Trong phần thực nghiệm, chúng tôi sẽ sử dụng độ đo *G-mean* để đánh giá hiệu quả phân lớp giữa các thuật toán.

2.4.2. Dữ liệu

Chúng tôi tiến hành thực nghiệm trên các bộ dữ liệu mất cân bằng của từ kho dữ liệu chuẩn quốc tế UCI [10]. Bảng 2 là thông tin về một số bộ dữ liệu mà bài báo sử dụng trong quá trình thực nghiệm.

Bảng 2. *Dữ liệu chuẩn quốc tế nguồn UCI*

Dữ liệu	**Số phần tử**	**Số thuộc tính**	**Tỉ lệ mất cân bằng**
Pima	768	8	1 : 2
Glass	193	9	1 : 6
Haberman	306	3	1 : 3
Blood	748	4	1 : 4
Breast-w	198	32	1 : 3

Các bộ dữ liệu trong bảng trên đều là các bộ dữ liệu có sự mất cân bằng lớp. Dữ liệu được gán nhãn hai lớp, lớp đa số được gán nhãn là negative và thiểu số được gán nhãn là positive. Trong đó, bộ dữ liệu Haberman và Breast-w có tỉ lệ mất cân bằng là 1:3; bộ dữ liệu Pima và bộ dữ liệu Blood có tỉ lệ mất cân bằng lần lượt là 1:2 và 1:4; bộ dữ liệu Glass có tỉ lệ mất cân bằng là 1:6.

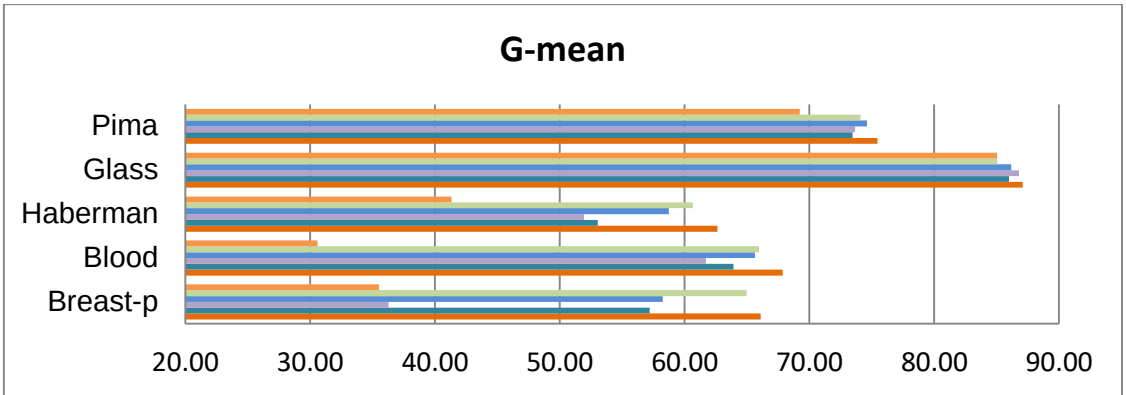
2.4.3. Kết quả thực nghiệm và đánh giá

Chúng tôi tiến hành thực nghiệm xây dựng trên ngôn ngữ R và Perl. Trong R sử dụng gói kernlab – package có chức năng phân lớp theo thuật toán SVM. Để đánh giá được hiệu quả phân lớp của hai thuật toán, chúng tôi kết hợp chúng với thuật toán phân lớp chuẩn SVM.

Đầu tiên chúng tôi chia ngẫu nhiên bộ dữ liệu ban đầu bằng phương pháp cross-validation ra làm 10-fold con có kích thước xấp xỉ nhau. Việc đánh giá thực hiện 10 lần, mỗi lần lấy một fold làm tập kiểm thử, 9 folds còn lại sử dụng làm tập huấn luyện. Với mỗi lần lặp, từ tập dữ liệu huấn luyện, chúng tôi thực hiện áp dụng một trong các thuật toán RO, RU, RSO, RSU, và RSOU để thu được tập dữ liệu huấn luyện mới. Tỷ lệ sinh phần tử nhân tạo và giảm phần tử được tính lần lượt dựa theo tham số *N%* và *M%*. Trong thực nghiệm, chúng tôi thực hiện với tham số *N%* lần lượt là 100%, 200%, 300%; tham số *M%* lần lượt là 5%, 10%, 15%, 20% và cuối cùng lựa chọn tham số tốt nhất cho từng dữ liệu.

Sau đó, áp dụng thuật toán phân lớp SVM vào bộ dữ liệu huấn luyện mới này thu được mô hình phân lớp để đưa vào đánh giá tập dữ liệu kiểm thử. Sau 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.

Sau khi cài đặt, thống kê kết quả, chúng tôi thực hiện đánh giá hiệu quả của các thuật toán trên từng bộ dữ liệu như Hình 1.



	Breast-p	Blood	Haberman	Glass	Pima
Original	35.50	30.59	41.33	85.03	69.22
RO	64.98	65.97	60.65	85.04	74.09
RU	58.26	65.64	58.74	86.18	74.62
RSO	36.27	61.70	51.94	86.80	73.66
RSU	57.21	63.92	53.05	86.00	73.47
RSOU	66.08	67.87	62.63	87.10	75.45

Hình 1. *Biểu đồ so sánh G-mean của các bộ dữ liệu UCI*

Bộ dữ liệu Pima với số phần tử là 768, khi áp dụng thuật toán mới RSOU (với tham số *N%* = 100%; *M%* = 5%) thì giá trị *G-mean* thu được là 75.45% cao hơn so với phương pháp sử dụng thuật toán RU có giá trị *G-mean* là 74.62%; phương pháp sử dụng thuật toán RSO có giá trị *G-mean* là 73.66%; phương pháp sử dụng thuật toán RSU có giá trị *G-mean* là 73.47% và phương pháp chỉ thực hiện phân lớp bộ dữ liệu gốc có giá trị *G-mean* là 69.22%.

Bộ dữ liệu Haberman với số phần tử là 306, khi áp dụng thuật toán mới RSOU (với tham số *N%* = 300%; *M%* = 10%) thì giá trị G-mean thu được là 62.63% cao hơn hẳn so với phương pháp sử dụng thuật toán RU có giá trị *G-mean* là 58.74%; phương pháp sử dụng thuật toán RSO có giá trị *G-mean* là 51.94%; phương pháp sử dụng thuật toán RSU có giá trị *G-mean* là 53.05% và phương pháp chỉ thực hiện phân lớp bộ dữ liệu gốc có giá trị *G-mean* là 41.33%.

Bộ dữ liệu Breast-p với số phần tử là 198, khi áp dụng thuật toán mới RSOU (với tham số *N%* = 100%; *M%* = 5%) thì giá trị *G-mean* thu được là 66.08% cao hơn hẳn so với phương pháp sử dụng thuật toán RU có giá trị *G-mean* là 58.26%; phương pháp sử dụng thuật toán RSO có giá trị *G-mean* là 36.27%; phương pháp sử dụng thuật toán RSU có giá trị *G-mean* là 57.21% và phương pháp chỉ thực hiện phân lớp bộ dữ liệu gốc có giá trị *G-mean* là 35.50%.

Biểu đồ trên so sánh hiệu quả phân lớp của các bộ dữ liệu bằng thuật toán Support Vector Machine (SVM) trước và sau khi điều chỉnh dữ liệu bởi RO, RU, RSO, RSU, và RSOU. Kết quả cho thấy, sau khi điều chỉnh bằng thuật toán RSOU, hiệu quả phân lớp có tăng lên, điển hình là bộ dữ liệu Breast-p, Blood, và Haberman tăng lên một cách đáng kể. Thuật toán RSOU sinh thêm phần tử mới dựa trên *k* láng giềng gần nhất của phần tử postive an toàn đồng thời xóa phần tử negative an toàn trong lớp đa số. Làm như vậy, RSOU không những làm giảm số phần tử lớp negative, mà còn làm tăng số phần tử positive một cách khoa học, tạo nên sự cân bằng dữ liệu, và nâng cao hiệu quả phân lớp dữ liệu.

3. KẾT LUẬN

Vấn đề mất cân bằng dữ liệu hiện nay đang rất được quan tâm vì ngày càng xuất hiện trong nhiều ứng dụng quan trọng trong thực tế như: y học, truyền thông, tài chính… Có nhiều hướng tiếp cận giải quyết vấn đề, trong bài báo này, chúng tôi đã trình bày tổng quan về thuật toán mới đề xuất dựa trên vùng an toàn nhằm nâng cao hiệu quả phân lớp dữ liệu. Thông qua việc nhân bản phần tử trong lớp thiểu số và giảm bớt phần tử trong lớp đa số dựa vào cấp độ an toàn của dữ liệu đã tạo ra khả năng khai phá những cơ sở dữ liệu có kích thước lớn bằng việc giảm mức độ mất cân bằng dữ liệu, đồng thời làm tăng độ chính xác, nâng cao hiệu quả tính toán của các kết quả phân lớp dữ liệu.

Trên cơ sở nghiên cứu và các kết quả thực nghiệm đạt được, chúng tôi nhận thấy có nhiều vấn đề cần được tiếp tục nghiên cứu. Đồng thời, chúng tôi sẽ nghiên cứu kết hợp

việc sinh thêm phần tử với các phương pháp khác như Boderline-SMOTE; Add-Boder-SMOTE đồng thời phát triển tiếp thuật toán RSOU để đạt được hiệu quả cao hơn trong việc giải quyết vấn đề mất cân bằng lớp.

TÀI LIỆU THAM KHẢO

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

[2] Y.-N. Zhang, D.-J. Yu, S.-S. Li, Y.-X. Fan, Y. Huang, and H.-B. Shen (2012). 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

[3] P. Xuan, M. Z. Guo, J. Wang, C. Y. Wang, X. Y. Liu, and Y. Liu (2011). Genetic algorithm-based efficient feature selection for classification of pre-miRNAs., *Genet. Mol. Res.*, vol. 10, no. 2, pp. 588–603.

[4] X. T. Dang, O. Hirose, T. Saethang, V. A. Tran, L. A. T. Nguyen, T. K. T. Le, M. Kubo, Y. Yamada, and K. Satou (2013). A novel over-sampling method and its application to miRNA prediction, *J. Biomed. Sci. Eng.*, vol. 6, no. 2, pp. 236–248.

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

[6] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, (2007). *Top 10 algorithms in data mining*, vol. 14, no. 1.

[7] H. Guo and H. L. Viktor (2004). "Learning from Imbalanced Data Sets with Boosting and Data Generation: The DataBoost-IM Approach," *SIGKDD Explor*. Newsl, vol. 6, no. 1, pp. 30–39.

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

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

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

Title: RANDOM-SAFE-OVER-UNDERSAMPLING - THE NOVEL METHOD BASED ON SAFE REGION TO IMPROVE PERFORMANCE OF IMBALANCED DATA CLASSIFICATION

Abstract: Researching on classification problem such as medical diagnosis, oil-overflowing incident discovery, economical and financial trick detection etc, is increasingly concerned by many scientists nowadays. However, many researchers have shown that the classification performance of these problems is not very high due to the difference among the number of elements between the data classes. One requirement is to have new approaches to deal with imbalanced data in order to increase the accuracy of the classification problems. Our paper proposes three novel methods based

on the safe level to enhance the effect of classification. Two methods, Random Safe Oversampling (RSO) and Random Safe Undersampling (RSU) are improved directly from Random Oversamling and Random Undersampling. The third method, Random Safe Over-Undersampling (RSOU), is a combination of RSO and RSU that simultaneously adds to the minority elements and removes the majority of elements in the safe regions.

Keywords: Imbalanced data; Classification; Safe level; Random Oversampling; Random Undersampling; Random Safe Oversampling; Random Safe Undersampling.