THUẬT TOÁN NEARMISS-SF CHO BÀI TOÁN PHÂN LỚP DỮ LIỆU MẤT CÂN BẰNG VÀ CÓ SỐ CHIỀU LỚN

NGUYỄN THỊ NGỌC HÀ[1,], NGUYỄN THỊ LAN ANH[2,**]*
[1]Học viên Cao học, Trường Đại học Sư phạm, Đại học Huế
[2]Khoa Tin học, Trường Đại học Sư phạm, Đại học Huế
[*]Email: hantn.quochochue@gmail.com
[**]Email: nguyenthilananh@dhsphue.edu.vn

Tóm tắt: Phân lớp dữ liệu mất cân bằng đặc biệt với tập dữ liệu có số chiều lớn là một bài toán quan trọng trong thực tế. Trong bài báo này chúng tôi đề xuất một thuật toán làm giảm số lượng phần tử lớp đa số trên một tập con các đặc trưng để cải thiện hiệu suất phân lớp tập dữ liệu mất cân bằng và có số chiều lớn.

Từ khóa: Dữ liệu mất cân bằng có số chiều lớn, phương pháp làm giảm số lượng phần tử lớp đa số, phương pháp lựa chọn đặc trưng.

1. ĐẶT VẤN ĐỀ

Bài toán phân lớp là một bài toán quan trọng trong học máy với nhiều thuật toán phân lớp khác nhau đã được phát triển và sử dụng rộng rãi trong nhiều lĩnh vực. Tuy nhiên trong thực tế có rất nhiều tập dữ liệu không cân bằng, là tập dữ liệu mà một lớp có số phần tử nhiều hơn các lớp khác theo tỷ lệ đáng kể, chẳng hạn là 1:100 hoặc 1:100000 [1]. Trong phạm vi bài toán phân lớp hai lớp của bài báo này, lớp có số lượng phần tử ít gọi là lớp thiểu số, lớp còn lại gọi là lớp đa số. Khi áp dụng các thuật toán phân lớp truyền thống lên các tập dữ liệu mất cân bằng, đa số các phần tử thuộc lớp đa số sẽ được phân lớp đúng và các phần tử thuộc lớp thiểu số cũng sẽ được gán nhãn lớp là nhãn lớp của lớp đa số. Điều này dẫn đến kết quả độ chính xác của việc phân lớp rất cao nhưng nhiều hoặc có khi là tất cả phần tử lớp thiểu số lại bị phân loại sai.

Có nhiều phương pháp đã được phát triển để xử lý bài toán phân lớp dữ liệu mất cân bằng [2][3][4][5][6]. Nhìn chung, các phương pháp để giải quyết vấn đề này được chia thành 2 nhóm: Các phương pháp tiếp cận ở mức độ dữ liệu và các phương pháp tiếp cận ở mức độ thuật toán. Một số nghiên cứu cho thấy các phương pháp tiếp cận ở mức dữ liệu hiệu quả hơn các phương pháp còn lại trong việc cải thiện độ chính xác sự phân lớp các tập dữ liệu mất cân bằng [1].

Các phương pháp tiếp cận ở mức dữ liệu là các phương pháp hướng đến thay đổi sự phân bố các đối tượng bằng cách sinh thêm các phần tử cho lớp thiểu số như SMOTE [2], và một số cải tiến của nó như Borderline-SMOTE [3], Save-Level-SMOTE [4],… hay giảm bớt các phần tử thuộc lớp đa số để làm giảm sự mất cân bằng giữa các lớp đối tượng như NEARMISS-1, NEARMISS-2, NEARMISS-3 [6],…

Tạp chí Khoa học, Trường Đại học Sư phạm, Đại học Huế
ISSN 1859-1612, Số 3(63)/2022: tr.130-136
Ngày nhận bài: 13/12/2021; Hoàn thành phản biện: 21/12/2021; Ngày nhận đăng: 22/12/2021

Xét về các phương pháp sinh thêm phần tử lớp thiểu số, SMOTE được cho là kỹ thuật được sử dụng thường xuyên nhất khi tập dữ liệu có số phần tử ít [2]. Thành công của SMOTE và các thuật toán cải tiến của nó chính là đơn giản, dễ tính toán, hiệu quả và hiệu suất vượt trội của chúng [2][3][4]. Tuy nhiên, khi áp dụng với dữ liệu có số chiều lớn, SMOTE và các thuật toán cải tiến của nó trở nên không phù hợp. Một trong những lý do là vì các thuật toán này sử dụng độ đo Euclid để tính khoảng cách giữa các phần tử. Các nghiên cứu đã chỉ ra rằng trong trường hợp này, các phần tử trong tập dữ liệu gần như là cách đều nhau. Ngoài ra, SMOTE sử dụng độ đo khoảng cách Euclid tức là giả định rằng tất cả các đặc trưng đều quan trọng như nhau trong việc xác định các phần tử láng giềng. Trong khi đó, tập dữ liệu nhiều chiều thường có các đặc trưng dư thừa làm giảm hiệu suất của thuật toán [5].

Để khắc phục vấn đề này, S. Maldonado và các cộng sự đã đề xuất thuật toán SMOTE-SF. SMOTE-SF xây dựng công thức khoảng cách mới dựa trên khoảng cách Minkowski [5] và chỉ có một tập con các đặc trưng $\boldsymbol{S_{Selected}}$ tham gia vào quá trình xác định các phần tử láng giềng của một phần tử lớp thiểu số X_i. Để tính khoảng cách giữa hai phần tử X_i và $X_{i'}$, SMOTE-SF sử dụng công thức:

$$\mathbf{d(X_i,X_{i'})} = (\sum_{j \in S_{Selected}} |X_{i,j} - X_{i',j}|^q)^{1/q} \quad (1)$$

Với $q \in \{1, 2, \infty\}$, lần lượt tương ứng khoảng cách Manhattan, Euclid và Chebyshev.

Tập con các đặc trưng $\mathbf{S_{Selected}}$ này được xác định dựa vào 4 phương pháp xếp hạng đặc trưng là Hệ số Fisher (Fisher Score), Thông tin tương hỗ (Mutual Information), Điểm tương quan (Correlation Score) và Độ trung tâm theo vector riêng (Eigenvector Centrality) [5].

Tuy nhiên, nếu bộ dữ liệu có số lượng phần tử lớn, việc sinh thêm phần tử sẽ khiến kích thước dữ liệu tăng lên, làm tăng chi phí thời gian và bộ nhớ cho giai đoạn phân lớp.

Trong khi đó, các phương pháp làm giảm số phần tử lớp đa số như Nearmiss-2[6] có thể làm giảm kích thước và thời gian huấn luyện bộ dữ liệu. Nearmiss-2 giải quyết vấn đề mất cân bằng dữ liệu bằng cách chọn giữ lại các phần tử lớp đa số gần với tất cả phần tử lớp thiểu số. Cụ thể với mỗi phần tử lớp đa số, 3 phần tử thuộc lớp thiểu số có khoảng cách xa nhất đến nó được xác định, sử dụng độ đo Euclid. Thuật toán tính giá trị trung bình khoảng cách đến 3 phần tử thiểu số xa nhất. Cuối cùng các phần tử thuộc lớp đa số có giá trị khoảng cách trung bình nhỏ nhất được chọn giữ lại. Tuy nhiên, phương pháp này có nhược điểm là dễ làm mất thông tin quan trọng của lớp đa số.

Trong bài báo này, chúng tôi đề xuất một phương pháp làm giảm số phần tử lớp đa số khắc phục các vấn đề được đề cập ở trên.

2. ĐỘ ĐO ĐÁNH GIÁ HIỆU SUẤT PHÂN LỚP

Đối với dữ liệu cân bằng, độ chính xác được sử dụng để đánh giá hiệu quả phân lớp. Nhưng đối với 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 phần tử lớp đa số được dự đoán đúng lớn, dẫn đến độ chính xác của mô hình phân lớp rất cao, trong khi rất ít hoặc hầu như không có

phần tử lớp thiểu số nào được dự đoán đúng. Vì vậy, các độ đo sau thường được sử dụng để đánh giá hiệu suất phân lớp trên các tập dữ liệu mất cân bằng [7].

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 TP_{rate} và TN_{rate}:

$$\textbf{G-mean} = \sqrt{TP_{rate} * TN_{rate}} \quad (2), \text{ với}$$

$$TP_{rate} = \frac{TP}{TP+FN}, \qquad TN_{rate} = \frac{TN}{TN+FP}$$

Ở đây, TP là số phần tử lớp thiểu số được dự đoán đúng; FP là số phần tử lớp đa số được dự đoán sai; TN là số phần tử lớp đa số được dự đoán đúng; FN là số phần tử lớp thiểu số bị dự đoán sai so với nhãn lớp thực của chúng.

Ngoài ra, **AUC** (*Area Under the ROC Curve):* đồ thị thể hiện hiệu suất của 1 mô hình phân lớp dựa trên 2 tham số FP_{rate} và TP_{rate}, cũng là 1 độ đo được sử dụng phổ biến trong nhiều bài báo khoa học. Đại lượng này chính là diện tích nằm dưới ROC curve. Giá trị này là một số dương nhỏ hơn hoặc bằng 1. Giá trị AUC càng lớn thì mô hình càng tốt.

3. PHƯƠNG PHÁP GIẢM PHẦN TỬ LỚP ĐA SỐ CHO TẬP DỮ LIỆU MẤT CÂN BẰNG VÀ CÓ SỐ CHIỀU LỚN NEARMISS-SF

Thuật toán NEARMISS-SF dưới đây được đề xuất để cải thiện hiệu suất bài toán phân lớp cho tập dữ liệu mất cân bằng và có số chiều lớn dựa trên cải tiến thuật toán SMOTE-SF và NEARMISS-2. Thuật toán **NEARMISS-SF** được mô tả như sau:

Algorithm *NEARMISS-SF*

Input: Tập đầy đủ các đặc trưng *S*; Tập các phần tử lớp thiểu số *T*; Tập các phần tử lớp đa số *N*; Tỉ lệ mẫu cần giữ lại *P*(%); Số lượng lân cận gần nhất *k*; Số lượng đặc trưng cần chọn *r*, giá trị *q* để tính khoảng cách.

Output: Tập các đặc trưng được chọn $S_{Selected}$; Tập các phần tử lớp đa số được giữ lại $N_{selected}$

1. $N_{selected} \leftarrow \{\}$
2. **for** $j \in S$
3. $FR(j) \leftarrow$ Giá trị Fisher Score của đặc trưng j.
4. **end for**
5. $S_{selected} \leftarrow$ r đặc trưng có giá trị Fisher Score lớn nhất
6. **for** $\mathbf{X}_i \in N$
7. **for** $\mathbf{X}_{i'} \in T$
8. $d(X_i,X_{i'}) = (\sum_{j \in S_{Selected}} |X_{i,j} - X_{i',j}|^q)^{1/q}$
9. **end for**
10. $TK \leftarrow$ Chọn k phần tử thiểu số xa X_i nhất và lưu vào mảng *TK*
11. $davg_i \leftarrow$ Trung bình cộng khoảng cách giữa X_i đến k phần tử trong *TK*

12. $cs_i \leftarrow i$ (* Lưu chỉ số của phần tử đa số đang xét*)

12. **end for**

13. *P ← P/100 * |N| (* Tính số lượng phần tử đa số cần giữ lại*)*

14. Sắp xếp davg tăng dần, cập nhật mảng cs tương ứng

15. *(*Kết nạp vào $N_{Selected}$ P phần tử lớp đa số có davg nhỏ nhất)*

16. **for** *i ←1* **to** *P*

17. $N_{Selected} \leftarrow \{N_{Selected}, X_{cs_i}\}$

18. **end for**

19. **end**

4. ĐÁNH GIÁ HIỆU SUẤT PHÂN LỚP

Chúng tôi thực hiện thực nghiệm trên các bộ dữ liệu UCI là Ecoli, Abalone, Yeast và Letter [8] để đánh giá hiệu suất của thuật toán. Thông tin về các bộ dữ liệu này được cho ở trong Bảng 1.

Bảng 1. *Các tập dữ liệu UCI*

Tập dữ liệu	Tỷ lệ mất cân bằng	Số lượng đặc trưng	Số lượng mẫu dữ liệu
Ecoli	8.6	7	336
Abalone	16.5	7	731
Yeast	28.1	8	1484
Letter	25.3	16	20000

Đối với thuật toán NEARMISS-2, NEARMISS-SF, giá trị k = 3 được chọn. Để giảm bớt sự mất cân bằng dữ liệu giữa các lớp đa số và và thiểu số, chúng tôi thực hiện thử nghiệm với các giá trị tham số P lần lượt bằng 4%, 5%, 10%, 20%, 30%, 40%. Đối với thuật toán SMOTE-SF, chúng tôi sử dụng giá trị N = 400, k = 5 như được đề xuất trong bài báo của S. Maldonado và cộng sự [5]. Các thuật toán SMOTE-SF, NEARMISS-SF sử dụng Hệ số Fisher làm phương pháp xếp hạng đặc trưng, các tham số q = {1, 2, ∞} và r nhận giá trị là một nửa số đặc trưng của tập dữ liệu. Thuật toán NEARMISS-2 vẫn sử dụng độ đo Euclid để tính khoảng cách.

Sau khi tiến hành điều chỉnh dữ liệu với các tham số như trên, chúng tôi sử dụng bộ phân lớp SMO cho SVM trong Weka [9] để tiến hành phân lớp và so sánh kết quả phân lớp trong trường hợp không có sự can thiệp của thuật toán làm thay đổi số phần tử để xử lý mất cân bằng dữ liệu (ORIGINAL), kết quả phân lớp có sử dụng thuật toán NEARMISS-2, kết quả phân lớp có sử dụng thuật toán SMOTE-SF, kết quả phân lớp có sử dụng thuật toán NEARMISS-SF nhằm đánh giá hiệu quả của thuật toán này.

Quá trình phân lớp được thực hiện như sau:

- Với mỗi tập dữ liệu, chúng tôi thực hiện mười lần 10-fold cross-validation, nghĩa là với mỗi lần thực hiện 10-fold cross-validation:

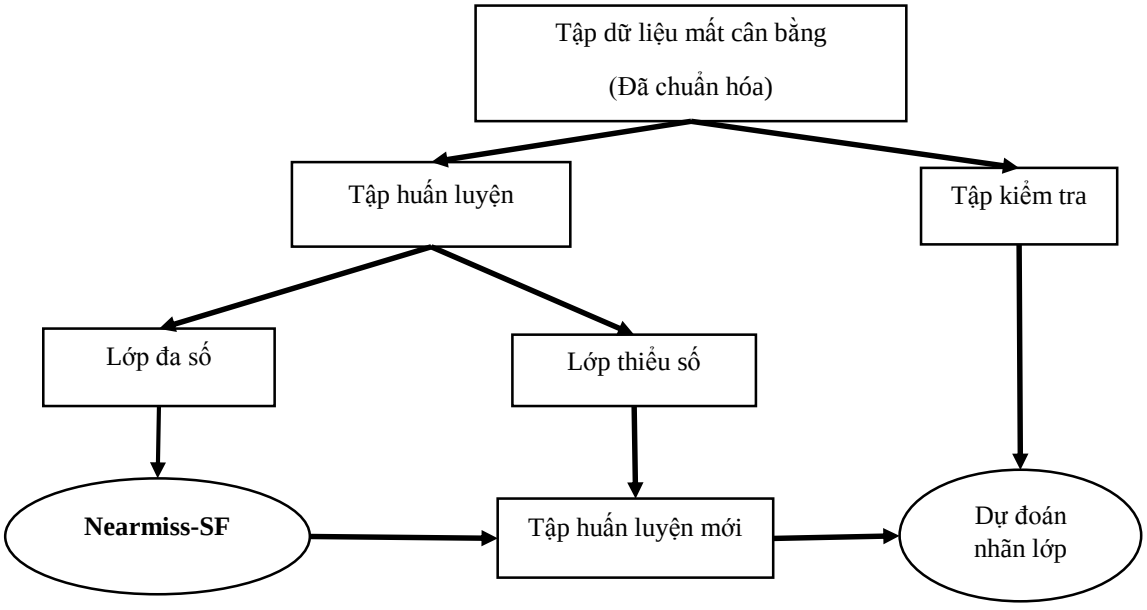
+ Tập dữ liệu được chia ngẫu nhiên thành 10 phần bằng nhau.

+ Lần lượt mỗi phần trong mười phần đó được chọn làm tập kiểm tra, chín phần còn lại tạo nên tập huấn luyện để xây dựng mô hình phân lớp.

+ Kết quả của thu được từ mười bộ tập kiểm tra và huấn luyện chính là kết quả của một lần thực hiện 10-fold cross-validation.

Cuối cùng, các giá trị độ đo đánh giá hiệu suất AUC và G-Mean được tính bằng cách lấy giá trị trung bình cộng của mười lần thực hiện độc lập này.

Toàn bộ quá trình phân lớp đánh giá hiệu suất thuật toán có áp dụng thuật toán Nearmiss-SF để xử lý vấn đề mất cân bằng dữ liệu được mô tả như ở Hình 1 bên dưới. Trường hợp sử dụng thuật toán Nearmiss-2, quá trình thực hiện tương tự như với trường hợp Nearmiss-SF. Đối với trường hợp sử dụng thuật toán SMOTE-SF, thay vì làm giảm số lượng phần tử lớp đa số, lớp thiểu số được tác động để làm tăng kích thước, các bước còn lại tương tự như trong Hình 1.



Hình 1. *Quá trình thực hiện phân lớp dữ liệu có áp dụng thuật toán Nearmiss-SF*

Kết quả đánh giá quá trình phân lớp của các bộ dữ liệu bằng các thuật toán khác nhau trong trường hợp tốt nhất được thể hiện trong Bảng 2, Bảng 3 và Bảng 4 bên dưới. Hai thuật toán Nearmiss-2 và Nearmiss-SF đạt AUC và G-mean tốt nhất với N = 20%, 40%, 5%, 4% tương ứng với các tập dữ liệu Ecoli, Abalone, Yeast, Letter. Thuật toán Nearmiss-SF và Smote-SF cho kết quả tốt nhất khi q = 1 hoặc ∞.

Kết quả này cho thấy, so với thuật toán Nearmiss-2, thuật toán Nearmiss-SF do chúng tôi đề xuất đã cải thiện hiệu quả phân lớp của toàn bộ các bộ dữ liệu được thử nghiệm. Điều

này cho thấy tác dụng của việc làm giảm số chiều của các bộ dữ liệu cũng như vai trò của việc lựa chọn độ đo khoảng cách trong thuật toán Nearmiss-SF.

So với thuật toán SMOTE-SF, thuật toán Nearmiss-SF cải thiện các độ đo AUC và G-mean trên ba tập dữ liệu Abalone, Yeast và Letter. Mặc dù AUC và G-mean của bộ dữ liệu Ecoli xấp xỉ với SMOTE-SF, giá trị TP_{Rate} của thuật toán do chúng tôi đề xuất lớn hơn gần 4% so với kết quả từ thuật toán SMOTE-SF.

Bảng 2. *Kết quả phân lớp theo độ đo AUC (%) của các tập dữ liệu UCI*

Tập dữ liệu	Original	Nearmiss-2	Smote-SF	Nearmiss-SF
Ecoli	50	76.39	**88.95**	88.51
Abalone	50	54.16	68.17	**69.32**
Yeast	50	70.36	71.36	**75.74**
Letter	85.36	84.96	91.87	**93.62**

Bảng 3. *Kết quả phân lớp theo độ đo TP_{Rate}(%),TN_{Rate} (%) của các tập dữ liệu UCI*

	TP_{Rate} (%)				TN_{Rate} (%)			
Tập dữ liệu	**Original**	**Nearmiss-2**	**Smote-SF**	**Nearmiss-SF**	**Original**	**Nearmiss-2**	**Smote-SF**	**Nearmiss-SF**
Ecoli	0	58.07	88.5	**92**	1.00	**93.75**	89.41	85.02
Abalone	0	37.43	36.60	**71.65**	1.00	73.49	**99.47**	67.00
Yeast	0	40.00	46.13	**58.47**	1.00	**97.90**	96.54	93.26
Letter	72.20	92.299	85.74	**95.86**	99.80	76.93	**98.04**	91.39

Bảng 4. *Kết quả phân lớp theo độ đo* **G-mean** *(%) của các tập dữ liệu UCI*

Tập dữ liệu	Original	Nearmiss-2	Smote-SF	Nearmiss-SF
Ecoli	0	68.17	**88.45**	88.02
Abalone	0	45.29	55.60	**68.35**
Yeast	0	63.93	63.03	**73.84**
Letter	84.09	84.56	91.66	**93.59**

5. KẾT LUẬN

Phân lớp dữ liệu mất cân bằng là một bài toán quan trọng và được ứng dụng vào nhiều lĩnh vực khác nhau trong thực tế. Một trong những kỹ thuật nâng cao hiệu suất của bài toán này là sử dụng phương pháp làm giảm phần tử của lớp đa số. Trong bài báo này, chúng tôi đã cải tiến thuật toán giảm phần tử lớp đa số NEARMISS-2 bằng cách kết hợp NEARMISS-2 với phương pháp lựa chọn đặc trưng để làm giảm tác động của sự nhiều chiều lên hiệu suất thuật toán. Chúng tôi tiến hành các thực nghiệm để so sánh, đánh giá hiệu suất của thuật toán trên bốn tập dữ liệu chuẩn UCI. Kết quả thực nghiệm đã cho thấy rằng thuật toán do chúng tôi đề xuất có hiệu quả trên bốn tập dữ liệu này dựa trên các giá trị độ đo đánh giá hiệu suất G-mean và AUC. Trong tương lai, có thể phát triển phương pháp này để áp dụng cho loại dữ liệu hỗn hợp (dữ liệu dạng số và phi số) bằng cách sử dụng một công thức khoảng cách khác như Gower [10].

TÀI LIỆU THAM KHẢO

[1] He H. and Garcia E. A. (2009). Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng*. 21 (9): 1263–1284.

[2] Chawla N. V, Bowyer K. W, Hall L. O., and Kegelmeyer W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* 16 (1): 321–357.

[3] Han H., Wang W.Y., and Mao B.H. (2005). *Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning*. Advances in Intelligent Computing, ICIC 2005, Lecture Notes in Computer Science. 3644

[4] Bunkhumpornpat C., Sinapiromsaran K., and Lursinsap C. (2009). *Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-sampling Technique for handling the class imbalanced problem*. PAKDD. 5476: 475–482

[5] Maldonado S., López J., Vairetti C. (2018). An alternative SMOTE oversampling strategy for high-dimensional datasets. *Applied Soft Computing Journal*. 76: 380–389

[6] Zhang Z. and Mani I. (2003). *KNN Approach to Unbalanced Data Distribution: A Case Study involving Information Extraction*. Workshop on Learning from Inmbalanced Datasets II, ICML, Washington DC

[7] Sun Y., Wong A. K. C., and Kamel M. S. (2009). Classification of Imbalanced Data: A Review. *Int. J. Pattern Recognit*. 23 (4): 687–719

[8] Lichman M. (2013). UCI Machine Learning Repository, *http://archive.ics.uci.edu/ml*, Irvine, CA: University of California, School of Information and Computer Science.

[9] Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Written L. (2009). *The WEKA Data Mining Software: An Update*. ACM SIGKDD Explorations Newsletter. 11 (1): 10-18.

[10] [10] Gower J.C. (1971). "A general coefficient of similarity and some of its properties". Biometrics. 27 (1): 857–874

Title: NEARMISS-SF ALGORITHM FOR DEALING WITH HIGH – DIMENSIONAL IMBALANCED DATA SETS

Abstract: Classifying the imbalanced data sets, especially the high-dimensional datasets is one of the important issues. In this paper, we present an undersampling algorithm with a subset of the available features to enhance the result of the high-dimensional imbalanced data sets classification.

Keywords: high-dimensional imbalanced data sets; undersampling methods; feature selection methods.