NÂNG CAO HIỆU SUẤT BÀI TOÁN PHÂN LỚP DỮ LIỆU MẤT CÂN BẰNG VỚI BLSMOTE-LOF

NGUYỄN THỊ PHƯƠNG NGA, NGUYỄN THỊ LAN ANH

KHOA TIN HỌC – TRƯỜNG ĐẠI HỌC SƯ PHẠM HUẾ

Tóm tắt: Phân lớp dữ liệu mất cân bằng hiện đang là một trong những vấn đề nhận được nhiều sự quan tâm của các nhà nghiên cứu trong nhiều lĩnh vực. Kỹ thuật Sinh thêm phần tử lớp thiểu số là một trong những phương pháp giúp cải thiện hiệu suất bài toán phân lớp dữ liệu mất cân bằng. Trong bài báo này, chúng tôi đề xuất một thuật toán sinh thêm phần tử thiểu số kết hợp với xử lí phần tử nhiễu để nâng cao kết quả bài toán phân lớp dữ liệu này.

Từ khóa: Dữ liệu mất cân bằng, sinh thêm phần tử lớp thiểu số, Smote-Lof, Boderline-Smote, hệ số ngoại lai cục bộ (Lof).

1. ĐẶT VẤN ĐỀ

Học máy từ dữ liệu mất cân bằng là một trong những vấn đề đã và đang được nhiều nhà nghiên cứu, trong nhiều lĩnh vực như bảo mật, viễn thông, tin sinh học… quan tâm. Một tập dữ liệu được gọi là mất cân bằng khi số lượng phần tử của một nhãn lớp nhỏ hơn nhiều so với số lượng phần tử của các nhãn lớp khác. Đối với bài toán phân lớp hai lớp, nhãn lớp có số phần tử nhiều hơn gọi là lớp đa số, lớp có số phần tử ít hơn gọi là lớp thiểu số. Dữ liệu mất cân bằng làm giảm chất lượng của việc phân lớp bằng các thuật toán phân lớp truyền thống vì thường kém chính xác trên lớp thiểu số. Điều này dẫn đến kết quả không mong muốn là không nhận dạng được những phần tử lớp thiểu số, thường là các đối tượng hiếm nhưng quan trọng trong thực tế[1]. Chẳng hạn như đối với bài toán chẩn đoán phát hiện sớm các bệnh nhân bị tiểu đường, việc phân lớp sai bệnh nhân bị tiểu đường thành người bình thường sẽ dẫn đến hậu quả nguy hiểm là người bệnh không được phát hiện và điều trị kịp thời.

Để giải quyết bài toán phân lớp dữ liệu mất cân bằng, có hai hướng tiếp cận chính, là dựa trên mức độ dữ liệu và dựa trên mức độ thuật toán. Sinh thêm phần tử lớp thiểu số và Giảm bớt phần tử lớp đa số là những phương pháp thường được áp dụng và được chứng minh là có hiệu quả. Một số phương pháp sinh thêm phần tử nhân tạo ở lớp thiểu số phổ biến như SMOTE[2], Borderline SMOTE[3], Safe-level SMOTE[4]...

SMOTE (Synthetic Minority Over-sampling Technique) sinh thêm các phần tử mới bằng cách: với mỗi phần tử x thuộc lớp thiểu số, chọn ngẫu nhiên một trong số k láng giềng

gần nhất cùng nhãn lớp của nó và lấy độ lệch giữa vector đặc trưng của x với láng giềng được chọn này nhân với một giá trị ngẫu nhiên trong đoạn [0,1] rồi cộng kết quả thu được với vectơ đặc trưng của x. Kết quả cuối cùng chính là vectơ đặc trưng của phần tử mới được sinh thêm của x [2].

Kỹ thuật SMOTE tạo ra phần tử sinh thêm nằm giữa phần tử lớp thiểu số được chọn và một trong các láng gềng của nó đã khắc phục được hạn chế của phương pháp Sinh thêm phần tử ngẫu nhiên, nhưng việc tạo ra các phần tử tổng hợp từ tất cả các phần tử thiểu số mà không quan tâm đến các phần tử lớp đa số dẫn đến ranh giới quyết định đối với lớp thiểu số lan rộng hơn vào không gian lớp đa số [5]. Năm 2002, Han H. và cộng sự đề xuất phương pháp Borderline-SMOTE [3]. Các phần tử lớp thiểu số được chia thành 3 nhóm: nhiễu, đường biên và an toàn bằng cách tính toán số phần tử thuộc lớp đa số trong m lân cận gần nhất. Khi đã xác định được các phần tử thiểu số ở đường biên, Borderline-SMOTE tiến hành sinh các phần tử mới tương tự như thuật toán SMOTE nhưng chỉ thực hiện với các phần tử nằm trên đường biên của lớp thiểu số thay vì tạo phần tử tổng hợp cho tất cả các phần tử của lớp thiểu số như SMOTE[3].

Một trong những nhược điểm khác của SMOTE là sinh ra nhiễu và các phần tử lớp thiểu số nhiễu được tạo ra này dễ bị nhận nhầm thành lớp đa số. Để khắc phục nhược điểm này, SMOTE – LOF [6], một phương pháp mở rộng của SMOTE kết hợp với Hệ số ngoại lai cục bộ (LOF) [7] để xử lý nhiễu trong lớp mất cân bằng đã được giới thiệu vào năm 2021. SMOTE-LOF xác định và loại bỏ nhiễu do SMOTE tạo ra bằng cách sử dụng hệ số ngoại lai cục bộ. Hệ số ngoại lai cục bộ được đề xuất để phát hiện ngoại lai dựa trên mật độ và hoàn toàn không có giả định về phân phối. Không giống như các phương pháp phát hiện ngoại lai truyền thống, coi ngoại lai là thuộc tính nhị phân, LOF chỉ định mức độ ngoại lai cho tất cả các phần tử dữ liệu [6].

Nhằm khắc phục các nhược điểm trên của SMOTE, trong bài báo này, chúng tôi đề xuất một phương pháp sinh thêm phần tử lớp thiểu số mới: thuật toán BLSMOTE-LOF.

Chúng tôi sẽ tiến hành thực nghiệm và đánh giá mô hình phân lớp trên một số tập dữ liệu sử dụng các thuật toán SMOTE – LOF, Borderline-SMOTE và thuật toán đề xuất BLSMOTE-LOF.

2. ĐỘ ĐO ĐÁNH GIÁ HIỆU SUẤT PHÂN LỚP

Để khắc phục nhược điểm của việc sử dụng độ chính xác đánh giá hiệu suất trong bài toán phân lớp dữ liệu mất cân bằng, chúng tôi sử dụng các độ đo sau:[8]

$$\text{Sensitivity} = \text{Recall} = \frac{TP}{TP+FN} \qquad (1)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \qquad (2)$$

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (3)$$

trong đó, 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.

Một độ đo khác thường được sử dụng trong bài toán có dữ liệu mất cân bằng là F-Measure. F-Measure là sự kết hợp giữa Precision và Recall:

$$\text{F-Measure} = \frac{2* \text{Precision} * \text{Recall}}{\text{Precision+Recall}} \qquad (4)$$

Ngoài ra, AUC (Area Under the ROC Curve)- đồ thị thể hiện hiệu suất của một mô hình phân lớp cũng thường được sử dụng. Đại lượng này chính là diện tích nằm dưới đường ROC. Giá trị AUC càng lớn thì mô hình càng tốt.

3. PHƯƠNG PHÁP BLSMOTE – LOF

Để cải thiện hiệu suất bài toán phân lớp dữ liệu mất cân bằng, chúng tôi đề xuất thuật toán BLSMOTE-LOF dưới đây. Thuật toán BLSMOTE-LOF được xây dựng dựa trên cải tiến thuật toán SMOTE- LOF và Borderline - SMOTE.

Thuật toán BLSMOTE-LOF

Input: Toàn bộ tập dữ liệu **T**, tập thiểu số **P** chứa các phần tử thuộc lớp thiểu số, tỷ lệ phần tử cần sinh thêm **S**%, số láng giềng gần nhất **m** của mỗi phần tử thiểu số để xác định phần tử thiểu số ở đường biên, số láng giềng gần nhất $\mathbf{k_S}$ để sinh phần tử tổng hợp, số lượng láng giềng gần nhất $\mathbf{k_{lof}}$ để xác định giá trị lof**.**

Output: Tập các phần tử sinh thêm **G**

1. O = Borderline-Smote(P,S,m,k_S)

2. Với mỗi phần tử x∈O:

    2.1. normalize(x)

    2.2. $N_k^{(x)}$ = knn(k)

    2.3. k_dist(x)

    2.4. reach-$dist_k$(x,o)

    2.5. lrd(x)

2.6. lof(x)

3. G ={ $x \in O | lof(x) < \varepsilon$}

Mô tả thuật toán:

1. Sinh thêm phần tử cho tập thiểu số P bằng thuật toán Boderline-SMOTE.

Các bước thực hiện của Borderline-SMOTE như sau:

Bước 1. Với mọi phần tử x thuộc lớp thiểu số P, xác định m' là số láng giềng gần nhất thuộc lớp đa số trong số m láng giềng gần nhất của x ($0 \leq m' \leq m$).

Bước 2.

Nếu m' = m nghĩa là m láng giềng gần nhất của x đều thuộc lớp đa số: x được xem là nhiễu (noise).

Nếu $m/2 \leq m' < m$, số lượng láng giềng gần nhất thuộc lớp đa số của x lớn hơn số lượng láng giềng gần nhất thuộc lớp thiểu số của nó: x là phần tử dễ bị dự đoán nhầm và được đưa vào tập DANGER.

Nếu $0 \leq m' < m/2$ thì x là phần tử an toàn.

Bước 3.

Với mỗi phần tử thuộc tập DANGER, xác định $\mathbf{k_S}$ láng giềng gần nhất thuộc lớp thiểu số P.

Bước 4. Sinh thêm $\mathbf{S} \times \mathbf{d_{num}}$ phần tử mới từ các phần tử thuộc tập DANGER, trong đó S là 1 số ngẫu nhiên từ 1 đến $\mathbf{k_S}$, $\mathbf{d_{num}}$ là số phần tử của tập DANGER.

Với mỗi $x \in P$, chọn ngẫu nhiên S phần tử trong số $\mathbf{k_S}$ láng giềng gần nhất của nó trong P, tính khoảng cách dif_j (j = 1,2,..., S) giữa x và S láng giềng gần nhất này.

Khi đó, S phần tử mới được sinh thêm của x là:

$$Syn(x) = \{synthetic_j \mid synthetic_j = x + r_j \times dif_j , j = 1,2,...,S\}$$

trong đó r_j (j = 1, 2,..., S) là một số ngẫu nhiên trong đoạn [0,1].

Tập tất cả các phần tử mới được sinh thêm O từ tập thiểu số P là O ={ Syn(x), $\forall x \in P, j$}

2. Tính hệ số ngoại lai cục bộ

2.1. Chuẩn hóa giá trị mỗi thuộc tính của phần tử lớp thiểu số

$$Z = \frac{X - \mu}{\sigma} \qquad (5)$$

trong đó,

μ: Giá trị trung bình của mỗi thuộc tính

σ: Độ lệch chuẩn của mỗi thuộc tính

2.2. Xác định k – láng giềng gần nhất cùng nhãn lớp của x: $N_k^{(x)} \subseteq o$ sao cho:

$$\forall o \in N_k(x), \forall y \in O, y \neq N_k(x) \Rightarrow dist(x, o) \leq dist(x, y) \quad (6)$$

trong đó,

O: Tập dữ liệu được sinh thêm bằng Borderline-SMOTE

k: số láng giềng gần nhất

dist (x, y): khoảng cách Euclide giữa x và $y \in O$

2.3. Tính giá trị k – distance(x): là khoảng cách lớn nhất trong số các khoảng cách từ phần tử lớp thiểu số đang xét x đến các láng giềng gần nhất của nó:

$$k\_dist(x) = \max\{dist(x,o) | o \in N_k(x)\} \quad (7)$$

2.4. Xác định khoảng cách có thể tiếp cận (reachability distance) giữa x và láng giềng của nó.

$$reach\_dist_k(x,o) = Max\{k\text{-}distance(o), dist(x, o) | o \in N_k(x)\} \quad (8)$$

với, k-distance(o): Giá trị k–distance của các đối tượng o trong tập tập các láng giềng gần nhất của x

2.5. Tính mật độ tiếp cận cục bộ của đối tượng x:

$$lrd(x) = \frac{k}{\sum_{o \in N_k(x)} reach\_dist(x,o)} \quad (9)$$

2.6. Tính hệ số ngoại lai cục bộ cuả x lof(x):

$$lof(x) = \frac{\sum_{o \in N_k(x)} \frac{lrd(o)}{lrd(x)}}{k} \quad (10)$$

3. Cuối cùng sau khi đã tính được hệ số ngoại lai cục bộ của các phần tử thiểu số từ thuật toán Borderline-SMOTE, nhiễu được xác định dựa vào hệ số ngoại lai cục bộ, là các phần tử có giá trị lof bé hơn ε, và loại khỏi tập phần tử được sinh thêm.

4. ĐÁNH GIÁ HIỆU SUẤT PHÂN LỚP

Để đánh giá hiệu suất của thuật toán BLSMOTE-LOF, chúng tôi sử dụng các bộ dữ liệu mất cân bằng UCI là Pima, Haberman và Breast-w [9]. Thông tin cụ thể về các bộ dữ liệu này, bao gồm số thuộc tính, số lượng phần tử trong mỗi tập dữ liệu, số lượng phần tử lớp thiểu số, số lượng phần tử lớp đa số và tỷ lệ mất cân bằng được mô tả ở Bảng 1.

Bảng 1. *Dữ liệu thực nghiệm*

Tập dữ liệu	**Số lượng thuộc tính**	**Số lượng mẫu**	**Số lượng lớp thiểu số**	**Số lượng lớp đa số**	**Tỷ lệ mất cân bằng IR**
Pima	9	768	268	500	1.87
Breast-w	11	695	241	454	1.88
Haberman	4	306	81	225	2.78

Đối với thuật toán Borderline-SMOTE và SMOTE, giá trị láng giềng gần nhất dùng để sinh thêm phần tử thiểu số là k = 5 được chọn. Điều chỉnh các tham số để tập dữ liệu sau khi đã Boderline-SMOTE, SMOTE có tỉ lệ giữa lớp đa số, thiểu số là 50:50 dựa trên phần đề xuất thực nghiệm trong bài báo của Asniar và cộng sự [6].

Đối với các thuật toán SMOTE-LOF và BLSMOTE-LOF, tham số k láng giềng gần nhất là k= 3 và k = 5 được chọn để thực nghiệm.

Thuật toán phân lớp SVM thuộc gói Caret[10] được sử dụng 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 SMOTE-LOF, kết quả phân lớp có sử dụng thuật toán Borderline-SMOTE, kết quả phân lớp có sử dụng thuật toán đề xuất BLSMOTE-LOF nhằm đánh giá hiệu quả của các thuật toán này.

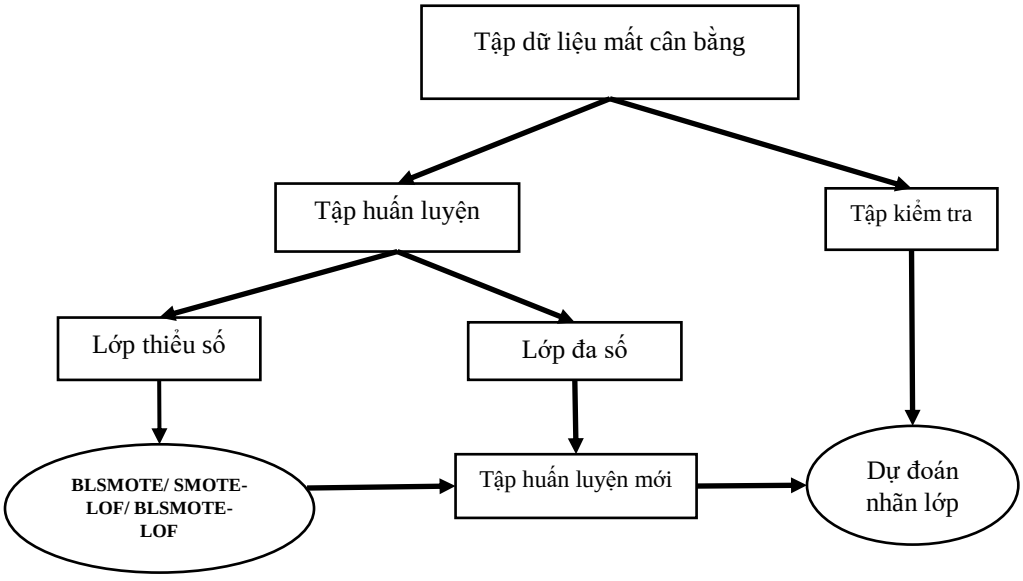
Quá trình phân lớp được thực hiện như sau: Với mỗi tập dữ liệu, chúng tôi thực hiện năm lần 5-fold cross-validation, nghĩa là với mỗi lần thực hiện 5-fold cross-validation:

+ Tập dữ liệu được chia ngẫu nhiên thành 5 phần bằng nhau.

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

+ Kết quả của thu được từ năm bộ tập kiểm tra và huấn luyện chính là kết quả của một lần thực hiện 5-fold cross-validation.

Cuối cùng, các giá trị độ đo đánh giá hiệu suất F-mesure và AUC được tính bằng cách lấy giá trị trung bình cộng của năm lần thực hiện độc lập này.
Toàn bộ quá trình thực nghiệm được mô tả như ở Hình 1.



Hình 1: *Quá trình thực nghiệm phân lớp dữ liệu*

Kết quả thực nghiệm trên các bộ dữ liệu Pima, Haberman và Breast-w lần lượt được trình bày ở Bảng 2, Bảng 3 và Bảng 4 tương ứng.

Kết quả thực nghiệm cho thấy, so với thuật toán Boderline-SMOTE và SMOTE – LOF thì thuật toán BLSMOTE-LOF có hiệu suất phân lớp tốt hơn hoặc tương đương. Thuật toán do chúng tôi đề xuất đã cải thiện hiệu quả phân lớp của các bộ dữ liệu được thử nghiệm. BLSMOTE-LOF đạt hiệu quả tốt nhất với giá trị k=5 trong xử lí nhiễu với phương pháp hệ số ngoại lai cục bộ LOF ở các giá trị (AUC, F-Measure).

Bảng 2. *Kết quả phân lớp của tập dữ liệu Pima*

Độ đo	**ORIGINAL**	**BODERLINE-SMOTE**	**SMOTE-LOF**		**BLSMOTE-LOF**	
			k=3	**k=5**	**k=3**	**k=5**
Sensitivity(%)	52.6	76.88	74.23	74.24	**77.45**	77.22
Specificity(%)	**88.8**	74.19	73.2	72.8	69.4	72.8
AUC(%)	70.7	74.14	73.71	73.52	73.43	**75.01**

F-Measure(%)	60.62	67.01	66.36	66.15	65.86	**67.85**

Bảng 3. *Kết quả phân lớp của tập dữ liệu Haberman*

Độ đo	**ORIGINAL**	**BODERLINE-SMOTE**	**SMOTE-LOF**		**BLSMOTE-LOF**	
			k=3	**k=5**	**k=3**	**k=5**
Sensitivity(%)	5.0	56.59	49.26	49.26	54.34	**54.34**
Specificity(%)	**98.22**	64.44	72.89	72.44	70.22	71.38
AUC(%)	51.61	60.57	61.08	60.86	62.1	**62.28**
F-Measure(%)	19.52	43.83	43.72	43.42	45.54	**45.6**

Bảng 4. Kết quả phân lớp của tập dữ liệu Breast-w

Độ đo	**ORIGINAL**	**BODERLINE-SMOTE**	**SMOTE-LOF**		**BLSMOTE-LOF**	
			k=3	**k=5**	**k=3**	**k=5**
Sensitivity(%)	98.76	99.17	98.76	98.76	**99.17**	**99.17**
Specificity(%)	**95.15**	95.15	95.37	95.15	**95.15**	**95.15**
AUC	96.95	97.16	97.06	96.95	**97.16**	**97.16**
F-Measure(%)	95.25	95.25	95.23	95.05	**95.25**	**95.25**

Cụ thể, BLSMOTE-LOF tốt hơn so với ORIGINAL, Boderline-SMOTE, SMOTE-LOF đối với tập dữ liệu Pima lần lượt là (4.31%, 7.23%), (0.87%,0.84%), (1.3%, 1.49%); đối với tập dữ liệu Haberman lần lượt là (10.67%, 26.08%), (1.71%, 1.77%), (1%, 1.88%); Đối với tập dữ liệu Breast-w thì các giá trị độ đo xấp xỉ nhau khi áp dụng các thuật toán xử lí mất cân bằng dữ liệu, giá trị AUC của BLSMOTE-LOF cải thiện hơn lần lượt là 0.21%, 0%, 0.1%.

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. Rất nhiều phương pháp khác nhau đã được đề xuất để cải thiện hiệu quả phân lớp loại dữ liệu này. Trong bài báo này, chúng tôi đã trình bày một phương pháp sinh thêm phần tử lớp thiểu số mới kết hợp giữa việc sinh thêm các phần tử ở đường biên và loại bỏ các phần tử mới được tạo thêm này mà không an toàn là

BlSmote-Lof. Chúng tôi cũng đã tiến hành các thực nghiệm trên một số tập dữ liệu UCI để đánh giá hiệu quả của thuật toán mới đề xuất. Kết quả thực nghiệm cho thấy rằng thuật toán BlSmote-Lof do chúng tôi đề xuất đã có những cải thiện tốt về hiệu suất phân lớp dữ liệu.

TÀI LIỆU THAM KHẢO

[1] Nguyễn Thị Lan Anh (2017), "Thuật toán HMU trong bài toán phân lớp dữ liệu mất cân bằng", *Tạp chí Khoa học và Giáo dục, Trường Đại học Sư phạm Huế ISSN 1859-1612*. Số 02(42), tr. 101.

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

[5] Zhuoyuan Zheng, Yunpeng Cai và Ye Li. (2015), "Oversampling Method for Imbalanced Classification", Computing and Informatics, Vol. 34, pp. 1017-1037.

[6] Asniar, Nur Ulfa Maulidevi, Karidanto Surendro. (2021), "SMOTE-LOF for noise identification in imbalanced data classification", Journal of King Saud University, pp.1-11.

[7] Markus M. Breunig (June 2000), "LOF: Identifying Density-Based Local Outliers.", *ACM SIGMOD Record* 2(29), pp. 93-104.

[8] Chawla N. V (2005), "Data Mining for Imbalanced Datasets: An Overview", *Data Mining and Knowledge Discovery Handbook.*

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

[10] https://CRAN.R-project.org/package=caret

Title: BLSMOTE-LOF ALGORITHM FOR DEALING WITH IMBALANCED DATA SETS

Abstract: Imbalanced classification is one of the most important research topics. Many approaches were developed to handle this problem. In this paper, we present a novel Oversampling algorithm for identifying the noises from the synthesized data to enhance the result of the imbalanced data sets classification.

Keywords: imbalanced data, Oversampling, Borderline-SMOTE, Local Outlier Factor

Tác giả:

NGUYỄN THỊ PHƯƠNG NGA

Trường Đại học Sư phạm, Đại học Huế

ĐT: 0962 856 054 - Email: nguyenthiphuongnga300886@gmail.com

NGUYỄN THỊ LAN ANH

Trường Đại học Sư phạm, Đại học Huế

ĐT: 0120 372 5257- Email: lananh257@gmail.com