Tạp chí Khoa học Đại học Cần Thơ
Phần A: Khoa học tự nhiên, Công nghệ và Môi trường
ISSN 1859-2333 | e-ISSN 2815-5599

DOI:10.22144/ctujos.2024.407

VẤN ĐỀ MẤT CÂN BẰNG DỮ LIỆU VÀ MỘT SỐ PHƯƠNG PHÁP XỬ LÝ DỮ LIỆU MẤT CÂN BẰNG TRONG MÔ HÌNH HỌC SÂU

Tống Lê Thanh Hải[1] và Phạm Ngọc Giàu[2,1*]
[1]Trường Đại học Tiền Giang, Việt Nam
[2]Trường Đại học Trà Vinh, Việt Nam
**Tác giả liên hệ (Corresponding author): phamngocgiau@tgu.edu.vn*

Thông tin chung (Article Information)

Nhận bài (Received): 31/01/2024
Sửa bài (Revised): 26/02/2024
Duyệt đăng (Accepted): 30/04/2024

Title: *The problem of data imbalances and some methods of processing imbalanced data in deep learning models*

Author(s): *Tong Le Thanh Hai[1] and Pham Ngoc Giau[2,1*]*

Affiliation(s): *[1]Tien Giang University, Viet Nam; [2]Tra Vinh University, Viet Nam*

TÓM TẮT

Trong bài viết này, vấn đề dữ liệu mất cân bằng, một hiện tượng phổ biến trong các bài toán phân loại nhị phân, khi mà số lượng mẫu của một lớp nhỏ hơn đáng kể so với lớp còn lại được đề cập đến. Nhiều phương pháp xử lý dữ liệu mất cân bằng trong học sâu được so sánh và đánh giá, bên cạnh đó sử dụng bộ dữ liệu Cat-Dog để nghiên cứu tác động của sự mất cân bằng đến quá trình phân loại. Các giải pháp được so sánh bao gồm cải tiến từ ba phương pháp tiếp cận: Data, Model và Loss, nhằm nâng cao hiệu suất dự đoán của các thuật toán máy học. Phương pháp tiếp cận Model qua việc áp dụng Transfer Learning với mô hình Resnet-18 cũng được đề xuất, đã được huấn luyện trước trên bộ dữ liệu ImageNet, cho kết quả F1-score là 95,19% và độ chính xác là 95,20% chỉ sau 10 epochs. Điều này cho thấy hiệu quả vượt trội so với các nghiên cứu trước đây tập trung vào cải thiện Data và Loss.

Từ khoá: *Dữ liệu mất cân bằng, phân loại nhị phân, tăng mẫu dữ liệu, giảm mẫu dữ liệu*

ABSTRACT

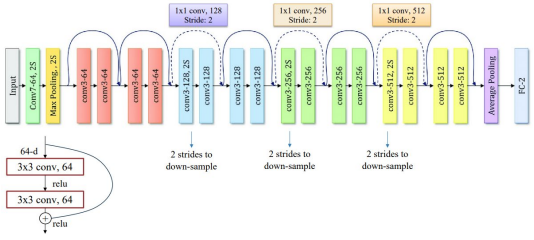
In this article, we address the problem of data imbalance, a common phenomenon in binary classification problems, where the sample number of one class is significantly smaller than the other. We compared and evaluated multiple approaches to processing imbalances in deep learning, using the Cat-Dog dataset to study the impact of imbalances on the classification process. The solutions compared include improvements from three approaches: Data, Model and Loss, aimed at enhancing the predictive performance of machine learning algorithms. We also recommend the Model approach by applying Transfer Learning with the Resnet-18 model, which was pre-trained on the ImageNet dataset, giving an F1-score of 95.19% and an accuracy of 95.20% after only 10 epochs. This showed superior efficacy compared to previous studies focused on improving data and loss.

Keywords: *Imbalanced Data, Binary Classification, Over-Sampling, Under-Sampling*

1. GIỚI THIỆU

Một tập dữ liệu được xem là mất cân bằng khi có sự chênh lệch lớn về số lượng mẫu giữa các lớp phân loại. Điều này xảy ra khi có một lớp chứa số lượng mẫu lớn hơn hẳn so với các lớp khác. Trong trường hợp này, một mô hình phân loại có thể đạt được độ chính xác cao khi dự đoán trên lớp có số lượng mẫu lớn, nhưng lại hoạt động kém hiệu quả trên lớp có ít mẫu. Do đặc điểm này, các thuật toán học máy truyền thống thường không phát huy hiệu quả trên tập dữ liệu mất cân bằng, dẫn đến việc phát triển các phương pháp và kỹ thuật mới nhằm cải thiện khả năng dự đoán của các thuật toán học máy trên những tập dữ liệu như vậy.

Bài viết này tập trung vào việc khám phá thuật toán lấy mẫu dữ liệu (data sampling) gồm có lấy mẫu giảm (under-sampling) và lấy mẫu tăng (over-sampling) và sự kết hợp giữa lấy mẫu giảm với lấy mẫu tăng (Duong & Dinh, 2023) nhằm giải quyết vấn đề mất cân bằng dữ liệu thông qua cách tiếp cận dựa trên dữ liệu. Ngoài ra, thuật toán Focal Loss (Lin et al., 2017) cũng được xem xét để giải quyết vấn đề tương tự nhưng thông qua cách tiếp cận dựa trên hàm mất mát và đề xuất giới thiệu phương pháp áp dụng Transfer Learning (Truong & Nguyen, 2022b) với mô hình Resnet-18 tiếp cận Model. Trong quá trình nghiên cứu, các thuật toán học máy được xây dựng, triển khai, thực nghiệm và đánh giá; sử dụng tập dữ liệu Cat-Dog (Bộ dữ liệu huấn luyện: 1000 ảnh mèo, 11000 ảnh chó; Bộ dữ liệu kiểm thử: 1500 ảnh mèo, 1500 ảnh chó). Thực nghiệm được thực hiện trên ba phương pháp tiếp cận khác nhau: Data, Model và Loss để giải quyết vấn đề dữ liệu mất cân bằng, áp dụng trên chín kỹ thuật khác nhau trong học máy, và đo lường hiệu suất qua chỉ số F1-score và accuracy. Kết quả nghiên cứu chỉ ra rằng việc lựa chọn thuật toán học máy phù hợp có tác động đáng kể đến hiệu quả của mô hình dự đoán.



Hình 1. Kiến trúc mô hình Resnet-18 (He, 2016)

2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Đặt vấn đề

Imbalanced Classification được hiểu là tình trạng một tập dữ liệu có sự phân phối không đồng đều giữa các lớp (class). Điều này thường gặp trong các bài toán phân loại nhị phân (Binary Classification), nơi mà số lượng mẫu thuộc một lớp cụ thể chỉ chiếm một tỷ lệ nhỏ so với lớp còn lại.

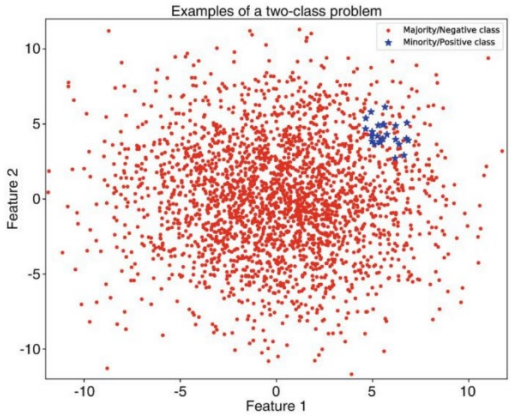
Ví dụ, xét bài toán dự đoán khách hàng có khả năng rời mạng của các công ty cung cấp gói cước mạng. Các công ty này thường chú trọng vào việc phát hiện khách hàng có xu hướng ngừng sử dụng dịch vụ để kịp thời triển khai các chiến lược chăm sóc và giữ chân họ, nhằm duy trì lợi nhuận. Trong tình huống này, tỷ lệ khách hàng ngừng sử dụng dịch vụ thường rất thấp so với số lượng khách hàng tiếp tục sử dụng (chỉ khoảng 5%). Dù 5% là một tỷ lệ nhỏ, nhưng nó lại có ý nghĩa lớn đối với lợi nhuận của công ty. Đối với bài toán phân loại mất cân bằng (imbalanced binary classification), tỷ lệ 5% và 95% tạo nên sự chênh lệch đáng kể, gây thách thức cho các kỹ sư AI trong quá trình lập kế hoạch và triển khai giải pháp.

Do sự mất cân bằng trong phân phối dữ liệu, hầu hết các thuật toán học máy không mang lại hiệu suất cao và cần được tinh chỉnh kỹ lưỡng để mô hình không chỉ dự đoán phổ biến (majority class) trong phần lớn trường hợp. Bên cạnh đó, một số chỉ số đánh giá như Accuracy không thể hiệu quả trong việc đánh giá các bài toán mất cân bằng.

Ví dụ, trong bài toán dự đoán khách hàng rời mạng (Duong & Dinh, 2023), nếu sử dụng Accuracy làm chỉ số đánh giá hiệu suất mô hình, nó tập trung vào độ chính xác trên cả hai lớp (rời mạng và không rời mạng). Tuy nhiên, thực tế cho thấy chúng ta chỉ quan tâm đến chất lượng của mô hình trên lớp khách hàng có khả năng rời mạng. Do đó, trong tình huống này, Accuracy không phải là chỉ số phù hợp để sử dụng trong đánh giá hoặc báo cáo.

Trong các bài toán dữ liệu không cân đối (Imbalanced Data), một thách thức phức tạp khác mà chúng ta thường gặp phải là việc hai lớp trong bài toán phân loại nhị phân, lớp có nhiều mẫu và lớp có ít mẫu thường có phân phối dữ liệu khác nhau. Điều này thường xảy ra trong các tình huống như phát hiện ngoại lệ (Outlier Detection), phát hiện bất thường (Anomaly Detection), chẩn đoán bệnh, hay phát hiện giao dịch đáng ngờ. Do đó, việc xây dựng một mô hình học máy có khả năng hiệu quả trên loại dữ liệu này cũng trở nên khó khăn.

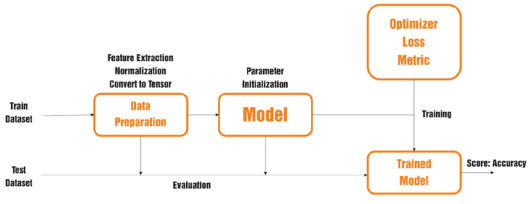
Dữ liệu không cân đối (Imbalance Data) tạo ra thách thức đáng kể trong việc mô hình hóa dự đoán (Predictive Modeling) bởi vì phần lớn các thuật toán phân loại được phát triển dựa trên giả định rằng số lượng mẫu trong mỗi lớp (Class) là tương đương nhau. Điều này thường dẫn đến việc các mô hình không đạt hiệu suất dự đoán tốt, đặc biệt là đối với các lớp có số lượng mẫu ít hơn (thiểu số). Trong bài viết này, ba phương pháp khác nhau được đề cập và nghiên cứu để xử lý vấn đề dữ liệu mất cân bằng, áp dụng trên chín kỹ thuật học máy khác nhau, nhằm tìm ra giải pháp hiệu quả cho vấn đề dữ liệu mất cân bằng.



Hình 2. Dữ liệu mất cân bằng giữa 2 lớp với tỉ lệ 1: 100 (Fernández et al., 2018b)

2.2. Phương pháp

Trong những năm gần đây, nhiều cách tiếp cận mới đã xuất hiện nhằm giải quyết vấn đề dữ liệu mất cân bằng trong phân lớp. Các nghiên cứu của Chawla et al. (2002), Paula et al. (2015), Buda et al. (2018a), Yu & Zhou (2021), Ghosh et al. (2022a) đã đóng góp đáng kể trong việc cải thiện hiệu quả của các ứng dụng đối mặt với vấn đề này. Mục tiêu của bài viết này là áp dụng ba phương pháp tiếp cận: Dữ liệu (Data), Mô hình (Model) và Hàm mất mát (Loss) để giải quyết những thách thức do dữ liệu mất cân bằng gây ra.



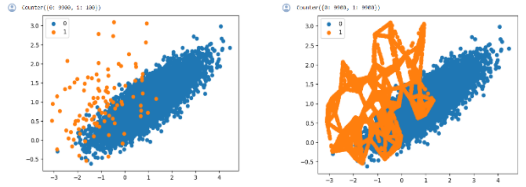
Hình 3. Mô hình được huấn luyện (Trained Model)

2.2.1. Tiếp cận Data

Lấy mẫu dữ liệu (Data Sampling) là quá trình tạo ra một phiên bản mới của tập dữ liệu, trong đó các điểm dữ liệu (Data Points) được chọn sao cho phân phối lớp (class) có sự khác biệt. Đây là một chiến lược đơn giản nhưng hiệu quả để xử lý vấn đề dữ liệu mất cân bằng. Có ba loại chính trong lấy mẫu dữ liệu: lấy mẫu giảm (Under-Sampling), lấy mẫu tăng (Over-Sampling) và kết hợp giữa lấy mẫu tăng và giảm (SMOTE – Synthetic Minority Over-Sampling Technique). Mục đích của những kỹ thuật này là giảm thiểu ảnh hưởng tiêu cực của việc phân bố lớp không cân đối trong quá trình huấn luyện mô hình.

Có nhiều phương pháp lấy mẫu dữ liệu (Data Sampling) được ưa chuộng và thường xuyên sử dụng, đặc biệt trong việc xử lý dữ liệu mất cân bằng. Dưới đây là một số phương pháp nổi bật:

• SMOTE (Synthetic Minority Over-sampling Technique) hoạt động bằng cách chọn các sample gần nhau trong feature space, vẽ một đường thẳng giữa các sample trong đó và tạo ra một sample mới tại một điểm nào đó trên đường thẳng đó. Một sample ngẫu nhiên từ minority class được chọn. Sau đó, tìm ra k sample gần nhất quanh sample đó (k = 5, tương tự KNN). Một neighbor được chọn ngẫu nhiên và một synthetic sample được tạo ra tại một điểm được chọn ngẫu nhiên giữa hai samples trong feature space (Chawla et al., 2002).



Hình 4. Tăng mẫu dữ liệu (Over Sampling)

• Under_Random (Random Under Sampling): kỹ thuật lấy mẫu giảm ngẫu nhiên là thực hiện việc loại bỏ một cách ngẫu nhiên một số lớp thuộc đa số lớp (Tahir et al., 2009).

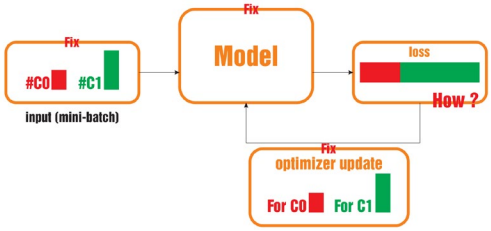
• Under_Clustering: kỹ thuật lấy mẫu giảm dựa vào gom cụm KMeans để thu giảm một số mẫu trong lớp đa số (Lin et al., 2018).

• Over_Aug (Over Augmentation): Kỹ thuật tăng cường dữ liệu và cải thiện khả năng của mô hình bằng cách tạo ra các biến thể mới từ dữ liệu hiện có thông qua các phép biến đổi như xoay, lật, thay đổi màu sắc…

• Over_Dup (Over Duplication): kỹ thuật tăng dữ liệu bằng cách tạo ra các mẫu trùng lặp.

Qua nghiên cứu và khảo sát tổng quan chúng tôi nhận thấy các kỹ thuật lấy mẫu tăng được sử dụng phổ biến hơn các kỹ thuật lấy mẫu giảm và khi số lượng mẫu của lớp thiểu số khá nhỏ so với số lượng của lớp đa số.

2.2.2. Tiếp cận Loss



Hình 5. Mất cân bằng tại hàm Loss

Phương pháp này chấp nhận dữ liệu đầu vào mất cân bằng tập trung vào việc đảm bảo rằng thông tin tín hiệu trong hàm mất mát (Loss Function) cần được cân bằng đúng cách. Để thực hiện điều này, các thuật toán sau đây được áp dụng nhằm tinh chỉnh và cân bằng giá trị trong hàm mất mát:

Class-Weighted Binary Cross-Entropy là một biến thể của hàm mất mát Binary Cross-Entropy (BCE) bằng việc thêm trọng số (W_i) cho các Class trong hàm mất mát để xử lý vấn đề mất cân bằng Class (Johnson et al., 2019).

Công thức của Class-weighted Binary Cross-Entropy ($BCE_{weighted}$):

$$BCE_{weighted} = -\frac{1}{N}\sum_{i=1}^{N}[w_1.y_i \log(p_i) + w_0.(1-y_i)\log(1-p_i)] \quad (1)$$

Trong đó:

N là số lượng mẫu trong tập dữ liệu

Y_i là nhãn thực tế của mẫu thứ i (thường là 0 hoặc 1)

P_i là xác suất dự đoán của mô hình cho mẫu thứ i

W_1 là trọng số cho lớp có nhãn 1

W_0 là trọng số cho lớp có nhãn 0

Cách chọn trọng số W_1 và W_0 phụ thuộc vào tỉ lệ mất cân bằng giữa các lớp trong tập dữ liệu. Một phương pháp phổ biến là đặt trọng số tỷ lệ nghịch với tần suất của mỗi lớp trong tập dữ liệu, giúp tăng ảnh hưởng của lớp ít xuất hiện hơn trong quá trình huấn luyện.

• Focal Loss là một biến thể của hàm mất mát Cross-Entropy được thiết kế để giảm tác động đến những Easy Samples mà tập trung vào các Hard Samples được giới thiệu bởi (Lin et al., 2017) trong bài báo "Focal Loss for Dense Object Detection".

Công thức Focal Loss (FL)

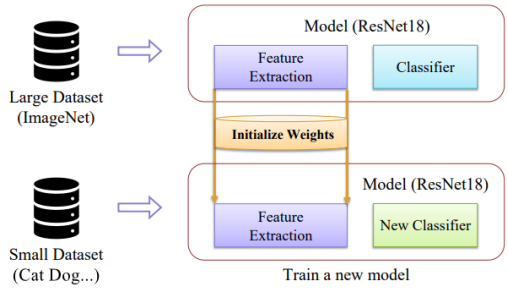
$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \quad (2)$$

P_t là xác suất dự đoán của mô hình cho lớp đúng. Nếu nhãn thực tế là 1, $p_t = p$ và nếu nhãn thực tế là 0, $p_t = 1 - p$ với p là xác suất dự đoán của mô hình cho lớp 1

α_t là trọng số dùng để cân bằng mất cân đối giữa các lớp.

γ là tham số focusting điều chỉnh mức độ mà các hàm mất mát tập trung vào Hard Samples và ít tập trung vào Easy Saples. Điều này giúp mô hình tập trung học từ những mẫu phức tạp hoặc ít xuất hiện hơn thay vì chỉ tập trung vào các mẫu dễ phân loại.

2.2.3. Tiếp cận Model



Hình 6. Học chuyển giao với mô hình Resnet-18

Phương pháp này tiếp nhận dữ liệu mất cân bằng và sử dụng kỹ thuật học chuyển giao (Transfer Learning) từ tập dữ liệu ImageNet, bao gồm 1.2 triệu hình ảnh thuộc 1000 danh mục khác nhau. Trong quá trình này, lớp cuối cùng của mô hình đã được đào tạo trên ImageNet được loại bỏ và thay thế nó bằng một hàm phân loại softmax mới, được thiết kế riêng cho 2 lớp có trong bộ dữ liệu thực tế đang khảo sát, đó là bộ dữ liệu Cat_Dog. Điều này giúp tận dụng kiến thức đã học được từ một tập dữ liệu lớn và phong phú, nhằm cải thiện hiệu suất trên bộ dữ liệu mục tiêu có tính chất mất cân bằng.

2.3. Độ đo đánh giá hiệu suất phân lớp

Trong trường hợp dữ liệu cân bằng, độ chính xác (Accuracy) thường được sử dụng để đánh giá hiệu quả phân lớp. Tuy nhiên, đối với dữ liệu mất cân bằng, việc dựa vào độ chính xác không còn hiệu quả. Lý do là trong dữ liệu mất cân bằng, lớp có số lượng

mẫu lớn (lớp đa số) thường được dự đoán chính xác hơn, làm cho độ chính xác tổng thể của mô hình trở nên cao một cách không thực tế, bất chấp việc hiệu suất dự đoán trên lớp thiểu số (có số lượng mẫu ít) có thể rất kém. Do đó, các độ đo sau thường được sử dụng để đánh giá hiệu suất phân lớp trên dữ liệu mất cân bằng (Johnson et al., 2019; Liu et al., 2020).

Confusion Matrix		Actual Label	
		Positive	Negative
Predicted Label	Positive	TP True Positive	FP False Positive
	Negative	FN False Negative	TN True Negative

True Positive (TP): Observation is positive, and is predicted to be positive
False Negative (FN): Observation is positive, but is predicted negative
True Negative (TN): Observation is negative, and is predicted to be negative
False Positive (FP): Observation is negative, but is predicted positive

Hình 7. Ma trận nhầm lẫn (Confusion Matrix) (Paula et al., 2015)

Độ chính xác (Accuracy) là tỷ lệ giữa số trường hợp được dự đoán chính xác True Positive và True Negative trên tổng trường hợp

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

Độ chính xác tuyệt đối (Precision) là tỷ lệ phần trăm các trường hợp True Positive dự đoán chính xác so với tổng số trường hợp dự đoán là Positive

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

Độ nhạy (Recall) là tỷ lệ phần trăm True Positive dự đoán chính xác so với tổng số trường hợp dự đoán là True Positive

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

$F_{measure}$ là một cách kết hợp giữa Precision và Recall thành một chỉ số duy nhất với β là trọng số giữa Precision và Recall

$$F_{Measure(\beta)} = \frac{(\beta^2 + 1) \times Precision \times Recall}{\beta^2 \times Precision \times Recall} \quad (6)$$

$F_{measure}$ khi β = 1 được gọi F1-score trở thành trung bình điều hoà của Precision và Recall

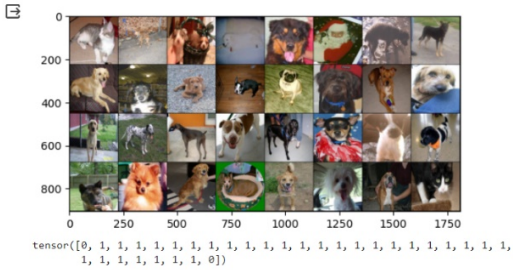
$$F1 - score = \frac{2 \times Precision \times Recall}{Precision \times Recall} \quad (7)$$

3. XÂY DỰNG VÀ HUẤN LUYỆN MÔ HÌNH

3.1. Chuẩn bị dữ liệu

Trong bài viết này, chúng tôi sử dụng bộ dữ liệu Cat_Dog Dataset, có sẵn trên thư viện Kaggle (https:/www.kaggle.com/). Bộ dữ liệu này bao gồm một tập dữ liệu huấn luyện (train) với 1000 hình ảnh mèo và 11000 hình ảnh chó, cùng với một tập dữ liệu kiểm thử (validation) gồm 1500 hình ảnh mèo và 1500 hình ảnh chó. Mỗi hình ảnh trong bộ dữ liệu này là một ảnh màu với kích thước khác nhau. Mỗi ảnh được gán nhãn từ 0 đến 1, phân loại vào 2 lớp, với số 0 tương ứng với lớp mèo và số 1 tương ứng với lớp chó.

Đồng thời, bộ dữ liệu ImageNet (https://www.image-net.org/) cũng được sử dụng, gồm 1,2 triệu hình ảnh thuộc 1000 danh mục khác nhau, để thực hiện việc huấn luyện (Pretraining) cho mô hình. Việc sử dụng dữ liệu từ ImageNet giúp cung cấp một nền tảng kiến thức rộng lớn cho mô hình, từ đó có thể tận dụng để cải thiện hiệu suất khi làm việc với bộ dữ liệu Cat_Dog.



Hình 8. Các mẫu dữ liệu trong Cat_Dog Dataset

3.2. Xây dựng và huấn luyện mô hình

Để theo dõi và lưu trữ giá trị mất mát (losses) và độ chính xác (accuracy) của mô hình trên tập huấn luyện (train set) và tập kiểm thử (validation set) sau mỗi kỳ huấn luyện (epoch), chúng tôi khởi tạo bốn danh sách (list): train_losses, train_acc, val_losses, và val_acc.

Train: Vòng lặp for chạy qua số lượng epoch đã xác định trước (num_epochs). Trong mỗi epoch: model được chuyển sang chế độ huấn luyện (model.train()).

– Một vòng lặp bên trong chạy qua tất cả các batch dữ liệu trong train_loader.

– Gradient của các tham số model được đặt về 0.

– Model tiến hành dự đoán và mất mát được tính toán.

– Gradient được tính toán thông qua phương thức backward().

– Các tham số model được cập nhật bằng thuật toán của optimizer.

– Giá trị mất mát và độ chính xác của model trên tập huấn luyện được log lại.

Evaluate: Sau khi huấn luyện xong trên tất cả các batch:

– Model được chuyển sang chế độ đánh giá (model.eval()).

– Test data được đưa qua model.

– Mất mát và độ chính xác trên tập test được log lại.

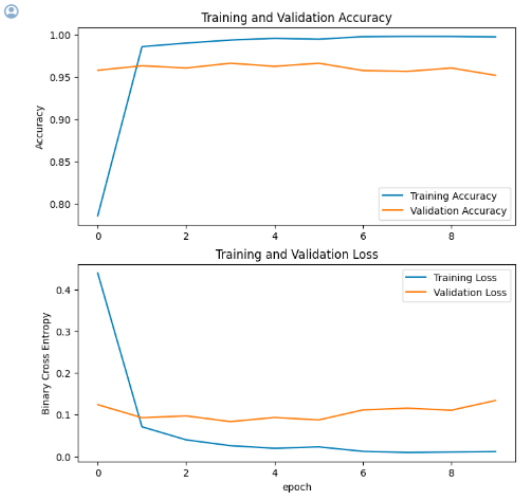
Lưu trữ và in kết quả: Giá trị mất mát trung bình và độ chính xác trung bình cho tập huấn luyện và tập kiểm thử được thêm vào các list tương ứng (xem hình 9).

```python
from sklearn.metrics import classification_report

target_names = ["Cat", "Dog"]

def cls_report(running_target, running_preds, target_names):
    targets = np.hstack(running_target[:-1])
    targets = np.concatenate((targets, running_target[-1]))

    predicts =  np.hstack(running_preds[:-1])
    predicts = np.concatenate((predicts, running_preds[-1]))

    report = classification_report(targets, predicts,
                                   target_names=target_names,
                                   zero_division=0.0, output_dict=True)
    del report['weighted avg']
    return report
```

Hình 9. Classification Report

4. KẾT QUẢ THỰC NGHIỆM

Để chuẩn bị và xử lý dữ liệu: hình ảnh đầu vào là ảnh màu với 3 kênh màu RGB sẽ được điều chỉnh kích thước thành 3x224x224 pixels để đảm bảo kích thước nhất quán cho mô hình, kích thước này thường được sử dụng trong các mô hình học sâu vì nó là một tiêu chuẩn không chính thức nhưng phổ biến cho hình ảnh đầu vào đặc biệt là trong các mô hình được huấn luyện trước như VGG, ResNet ... sau đó chuyển hình ảnh thành tensor PyTorch và chuẩn hoá dữ liệu bằng cách sử dụng giá trị trung bình (mean) và độ lệch chuẩn (std) là 0,5 cho mỗi kênh màu, đưa giá trị pixel về phạm vi [-1, 1] giúp cải thiện quá trình huấn luyện. Mô hình sử dụng là kiến trúc ResNet18 (cụ thể là models.resnet18 với weights=None), trong đó lớp cuối cùng được loại bỏ và thay thế bằng một lớp mới phù hợp với hai lớp của dữ liệu đích: Mèo và Chó. Đối với quá trình tối ưu hóa, mô hình sẽ sử dụng thuật toán Adam làm optimizer và Cross Entropy Loss làm hàm mất mát, với tốc độ học được thiết lập là 0,0001. Thực nghiệm được thiết lập để kiểm tra chín kỹ thuật khác nhau, áp dụng ba phương pháp tiếp cận: Data, Model và Loss, nhằm giải quyết vấn đề dữ liệu mất cân bằng.



Hình 10. Độ chính xác và độ mất mát trong quá trình huấn luyện sử dụng phương pháp học chuyển giao

Bảng 1. Kết quả thực nghiệm các phương pháp tiếp cận Data, Model và Loss

METHOD	EPOCH		METRIC	MACRO AVG	CAT	DOG
Over_Aug	50	Train		Loss: 0,87%, Accuracy: 99,68%		
			Precision	99,68%	99,62%	99,62%
			Recall	99,68%	99,74%	99,74%
			F1-score	99,68%	99,68%	99,68%
		Valid		Loss: 127,69%, Accuracy: 79,05%		
			Precision	82,04%	91,91%	91,91%
			Recall	79,00%	63,60%	63,60%
			F1-score	78,49%	75,18%	75,18%
Over_Dup	50	Train		Loss: 2,53%, Accuracy: 99,12%		
			Precision	99,12%	99,09%	99,09%
			Recall	99,12%	99,15%	99,15%
			F1-score	99,12%	99,12%	99,12%

METHOD	EPOCH		METRIC	MACRO AVG	CAT	DOG
		Valid		Loss: 217,86%, Accuracy: 63,26%		
			Precision	72,53%	87,05%	87,05%
			Recall	63,17%	30,93%	30,93%
			F1-score	58,90%	45,65%	45,65%
Over_SMOTE	50	Train		Loss: 0,02%, Accuracy: 100%		
			Precision	100%	100%	100%
			Recall	100%	100%	100%
			F1-score	100%	100%	100%
		Valid		Loss: 142,84%, Accuracy: 69,76%		
			Precision	71,09%	76,43%	76,43%
			Recall	69,73%	57,07%	57,07%
			F1-score	69,24%	65,34%	65,34%
Under_Clustering	50	Train		Loss: 7,07%, Accuracy: 97,36%		
			Precision	97,45%	97,59%	97,59%
			Recall	97,45%	97,30%	97,30%
			F1-score	97,45%	97,45%	97,45%
		Valid		Loss: 253,87%, Accuracy: 61,08%		
			Precision	68,32%	79,89%	79,89%
			Recall	61,00%	29,40%	29,40%
			F1-score	56,67%	42,98%	42,98%
Under_Random	50	Train		Loss: 1,97%, Accuracy: 99,56%		
			Precision	99,70%	99,40%	99,40%
			Recall	99,70%	100%	100%
			F1-score	99,70%	99,70%	99,70%
		Valid		Loss: 136,97%, Accuracy: 64,33%		
			Precision	67,91%	59,94%	59,94%
			Recall	64,37%	86,60%	86,60%
			F1-score	62,51%	70,85%	70,85%
Resnet-18 (weight=none)	20	Train		Loss: 0,18%, Accuracy: 99,95%		
			Precision	99,88%	99,80%	99,96%
			Recall	99,79%	99,60%	99,98%
			F1-score	99,84%	99,70%	99,97%
		Valid		Loss: 246,88%, Accuracy: 62,90%		
			Precision	75,82%	94,08%	57,56%
			Recall	62,90%	27,53%	98,27%
			F1-score	57,60%	42,60%	72,59%
Class weight	20	Train		Loss: 0,19%, Accuracy: 99,98%		
			Precision	99,85%	99,70%	100%
			Recall	99,99%	100%	99,97%
			F1-score	99,92%	99,85%	99,99%
		Valid		Loss: 153,87%, Accuracy: 69,27%		
			Precision	75,33%	87,73%	62,94%
			Recall	69,27%	44,80%	93,73%
			F1-score	67,31%	59,31%	75,31%
Focal Loss	20	Train		Loss: 0%, Accuracy: 100%		
			Precision	100%	100%	100%
			Recall	100%	100%	100%
			F1-score	100%	100%	100%
		Valid		Loss: 39,00%, Accuracy: 59,23%		
			Precision	74,29%	93,42%	55,17%
			Recall	59,23%	19,87%	98,60%

METHOD	EPOCH		METRIC	MACRO AVG	CAT	DOG
			F1-score	51,76%	32,77%	70,75%
Transfer Learning	10	Train		Loss: 1,16%, Accuracy: 99,74%		
			Precision	99,05%	99,21%	99,88%
			Recall	99,27%	98,70%	99,84%
			F1-score	99,16%	98,45%	99,86%
		Valid		Loss: 13,39%, Accuracy: 95,20%		
			Precision	95,61%	99,93%	91,29%
			Recall	95,20%	90,47%	99,93%
			F1-score	95,19%	94,96%	95,42%

Kết quả thực nghiệm đã cho thấy hầu hết các phương pháp đều đạt hiệu suất cao trong giai đoạn huấn luyện, với độ chính xác thường đạt gần hoặc bằng 100% cho các phương pháp như Over_SMOTE và Focal Loss. Hiệu suất cao này cho thấy các mô hình có khả năng học sự khác biệt giữa mèo và chó từ dữ liệu huấn luyện.

Biến động hiệu suất tập kiểm thử: Tuy nhiên, hiệu suất tập kiểm thử lại cho thấy sự biến động đáng kể giữa các phương pháp khác nhau. Chẳng hạn, trong khi phương pháp Mô hình tiền huấn luyện mang lại độ chính xác cao nhất là 95.20% trên dữ liệu tập kiểm thử, phương pháp Focal Loss lại chỉ đạt độ chính xác thấp là 59.23%, cho thấy sự chênh lệch trong khả năng tổng quát hóa của các phương pháp khác nhau đối với dữ liệu chưa từng thấy.

Dấu hiệu của hiện tượng quá khớp: Sự giảm sút đáng kể trong hiệu suất từ huấn luyện sang kiểm thử cho một số phương pháp (ví dụ, Over_Aug, Over_Dup) có thể chỉ ra hiện tượng quá mức khớp, nơi mô hình học quá tốt dữ liệu huấn luyện, bao gồm cả nhiễu và dữ liệu ngoại lai, dẫn đến hiệu suất kém trên dữ liệu mới, chưa từng thấy.

Hiệu quả của mô hình học chuyển giao: Phương pháp Transfer Learning với mô hình Resnet-18 nổi bật với độ chính xác đánh giá cao nhất 95.20%, cho thấy việc sử pretrain model và tinh chỉnh chúng cho các nhiệm vụ cụ thể có thể cải thiện đáng kể vấn đề mất cân bằng dữ liệu.

5. KẾT LUẬN

Bài viết này đưa ra góc nhìn sâu sắc về một thách thức lớn trong học máy và học sâu, đó là vấn đề dữ liệu mất cân bằng. Điều này đặc biệt quan trọng trong các tình huống phát hiện ngoại lệ, nơi mà dữ liệu cần dự đoán có xu hướng bị át đi bởi các dữ liệu khác chiếm đa số.

Bài viết khám phá và đánh giá một loạt các phương pháp tiếp cận từ việc điều chỉnh tập dữ liệu, tinh chỉnh mô hình, đến tối ưu hóa hàm mất mát để giải quyết vấn đề này. Nổi bật trong số đó là sự hiệu quả của kỹ thuật Transfer Learning khi áp dụng với mô hình Resnet-18, mang lại một giải pháp mạnh mẽ và đa dạng để nâng cao hiệu suất mô hình trên dữ liệu mất cân bằng.

Bài viết cung cấp một cái nhìn toàn diện và chi tiết về các phương pháp tiếp cận và giải quyết dữ liệu mất cân bằng trong lĩnh vực học sâu, đóng góp một hướng tiếp cận mới cho cộng đồng nghiên cứu và ứng dụng học máy.

TÀI LIỆU THAM KHẢO (REFERENCES)

Chawla, N. V, Bowyer, K. W, Hall, L. O., & Kegelmeyer, W. P. (2002). *SMOTE : Synthetic Minority Over-sampling Technique*. J. Artif. Intell. Res., *16*(1), 321–357. https://doi.org/10.1613/jair.953

Tahir, M. A., Kittler, J., Mikolajczyk, K., & Yan, F. (2009). *A multiple expert approach to class imbalance problem using inverse random undersampling.* Proc. Of Int. Workshop on Multiple Classifier Systems, pp. 82-91. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-02326-2_9

Paula, B., Torgo, L., & Ribeiro, R. (2015). *A survey of predictive modelling under imbalanced distributions*. arXiv preprint arXiv, 1505(01658). https://doi.org/10.1109/ICCV.2017.324

He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition.* In Proceedings of the IEEE conference on computer vision and pattern recognition, 770-778. https://doi.org/10.1109/CVPR.2016.90

Lin, T. Y., Goyal, P., Girshick, R., He, K., Dollar, P. (2017). *Focal loss for dense object selection*, Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2980-2988. https://doi.org/10.1109/ICCV.2017.324

Buda, M., Maki, A., & Mazurowski, M. A. (2018a). A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106, 249-259. https://doi.org/10.1016/j.neunet.2018.07.011

Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., & Herrera, F., (2018b). *Learning from Imbalanced Data Sets, Learning from Imbalanced Data Sets*. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-98074-4

Johnson, J., & Khoshgoftaar, T. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, 6, 27. https://doi.org/10.1186/s40537-019-0192-5

Liu, Z., Cao, W., Gao, Z., Bian, J., & Chen, H. (2020). Self-paced Ensemble for Highly Imbalanced Massive Data Classification. *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, Dallas, TX, USA, 841-852. https://doi.org/10.1109/ICDE48307.2020.00078

Yu, L., & Zhou, N. (2021). *Survey of imbalanced data methodologies*. https://doi.org/10.48550/arXiv.2104.02240

Ghosh, K., Bellinger, C., Corizzo, R., Japkowicz, N., Branco, P., & Krawczyk, B. (2022a). *The class imbalance problem in deep learning. Mach Learn*. https://doi.org/10.1007/s10994-022-06268-8

Thanh, T. T. P., & Nghe, N. T. (2022b). Rice Leaf Disease Recognition Using Transfer Learning Method. *Can Tho University Journal of Science*, 58(4), 1-7 *(in Vietnamese)*. https://doi.org/10.22144/ctu.jvn.2022.157

Duong, T. A., & Dinh, M. H. (2023). Classifying Imbalanced Data in Customer Churn Prediction Using an Improved Random Forest Algorithm. *HUFLIT Journal of Science*, *7*(3), 58-58. *(in Vietnamese)* https://hjs.huflit.edu.vn/index.php/hjs/article/view/143

ImageNet web page. https://www.image-net.org/

Kaggle web page. https:/www.kaggle.com/