

DỰ ĐOÁN KHẢ NĂNG THI HỎNG MÔN CỦA HỌC SINH BẰNG MẠNG ĐỐI KHÁNG TẠO SINH CÓ ĐIỀU KIỆN CT-TABGAN

PREDICTING THE LIKELIHOOD OF STUDENTS TO FAIL EXAM USING THE CONDITIONAL GENERATIVE RESISTANCE NETWORK CT-TABGAN

Nguyễn Thị Lan Anh, Vĩnh Anh Nghiêm Quân*

* Trường ĐH Sư Phạm - Đại học Huế

Received: 02/04/2026; Accepted: 15/05/2026; Published: 15/06/2026

Tóm tắt: Việc dự báo sớm nguy cơ trượt kỳ thi của sinh viên đóng vai trò quan trọng, giúp các cơ sở giáo dục triển khai kịp thời các chiến lược tư vấn và can thiệp, từ đó cải thiện kết quả học tập. Mặc dù nhiều nghiên cứu đã ứng dụng các kỹ thuật học máy để dự báo kết quả học tập và phân tích các yếu tố ảnh hưởng đến thành tích của sinh viên, hiệu quả dự báo vẫn bị hạn chế bởi vấn đề mất cân bằng dữ liệu. Trong bối cảnh đó, các mô hình phân loại thường có xu hướng thiên lệch về phía lớp đa số, dẫn đến việc phân loại sai và bỏ sót những sinh viên thực sự có nguy cơ trượt. Để giải quyết thách thức này, nhiều kỹ thuật xử lý mất cân bằng dữ liệu đã được đề xuất ở cả cấp độ thuật toán lẫn dữ liệu; trong đó, các mô hình sinh dữ liệu dựa trên GAN đã cho thấy tiềm năng lớn nhờ khả năng tạo ra dữ liệu tổng hợp chất lượng cao. Nghiên cứu này sử dụng kiến trúc CT-TabGAN – một biến thể của GAN được thiết kế chuyên biệt cho dữ liệu dạng bảng – nhằm nâng cao độ chính xác trong việc dự báo kết quả học tập của sinh viên trong điều kiện dữ liệu bị mất cân bằng.

Từ khóa: dự báo kết quả học tập của sinh viên, CT-TabGAN, dữ liệu mất cân bằng

Abstract: Early prediction of student exam failure plays a crucial role in enabling educational institutions to implement timely counseling and intervention strategies, thereby improving academic outcomes. Although numerous studies have employed machine learning techniques to predict student performance and analyze the factors influencing academic achievement, predictive effectiveness remains constrained by the issue of data imbalance. In such contexts, classification models tend to be biased toward the majority class, leading to the misclassification and overlooking of students who are genuinely at risk of failure. To address this challenge, various data imbalance handling techniques have been proposed at both the algorithmic and data levels, among which GAN-based data generation models have demonstrated considerable potential due to their ability to produce high-quality synthetic data. In this study, the CT-TabGAN architecture, a GAN variant specifically designed for tabular data, is employed to enhance the accuracy of student academic performance prediction under imbalanced data conditions.

Keywords: student performance prediction, CT-TabGAN, imbalanced data

1. Đặt vấn đề

Nhận diện và dự báo sớm nguy cơ hỏng môn của học sinh đóng vai trò quan trọng giúp nhà trường kịp thời đưa ra các giải pháp tư vấn, hỗ trợ hiệu quả, giảm thiểu rủi ro học tập cho học sinh (HS), vì điểm thi cuối kỳ của mỗi môn học phản ánh trực tiếp kết quả học tập cả năm và định hình lộ trình phát triển tương lai của người học. Đây là một trong những bài toán nhận được sự quan tâm lớn của các nhà nghiên cứu.

Nhiều chương trình dự đoán kết quả học tập của HS cũng như để phân tích các yếu tố ảnh hưởng đến kết quả này đã được thực hiện, ứng dụng các công cụ học máy [1][2]. Tuy nhiên, sự mất cân bằng dữ liệu tồn tại trong các tập dữ liệu là một trong những nguyên nhân khiến kết quả dự đoán vẫn còn hạn chế. Khi dữ liệu bị mất cân bằng, các bộ phân lớp này có khuynh hướng nhận dạng tất cả các đối tượng đều thuộc về lớp các đối tượng có số lượng lớn. Khi áp dụng vào bài toán dự đoán kết quả thi của học sinh,

những HS có khả năng thi hỏng lại được dự đoán là đạt, nên không được tư vấn kịp thời để cải thiện kết quả, có khả năng dẫn đến những điều đáng tiếc.

Để xử lý vấn đề mất cân bằng dữ liệu trong bài toán phân lớp, nhiều kỹ thuật khác nhau đã được phát triển, bao gồm các phương pháp tiếp cận ở mức độ thuật toán như điều chỉnh xác suất ước lượng, sử dụng các hằng số phạt khác nhau cho các nhãn lớp khác nhau; và các phương pháp tiếp cận ở mức dữ liệu như sinh thêm các phần tử cho lớp thiểu số, giảm bớt các phần tử thuộc lớp đa số; hoặc các phương pháp kết hợp [3]. Phương pháp sinh thêm phần tử GAN có ưu điểm là khả năng sinh dữ liệu có độ chân thực và chất lượng cực cao, học và bảo toàn được các mối quan hệ phức tạp cùng với khả năng tạo dữ liệu nhanh chóng. Tuy nhiên, GAN lại khó áp dụng cho dữ liệu dạng bảng, vì các đặc điểm của nó. Để tận dụng được ưu điểm của GANs, CT-TabGAN đã được phát triển và cho kết quả tốt. Do vậy, trong nghiên cứu này, chúng tôi sử dụng kiến trúc CT-TabGAN

cho bài toán dự đoán khả năng thi hỏng của HS.

2. Nội dung nghiên cứu

2.1 Mô hình CT-TabGAN

Mạng đối kháng tạo sinh GAN (Generative Adversarial Networks) là một kiến trúc trí tuệ nhân tạo học sâu, chuyên dùng để tạo ra dữ liệu mới, được ứng dụng rộng rãi trong xử lý hình ảnh, video và âm thanh. Cấu trúc của GAN bao gồm hai mạng nơ-ron hoạt động đối kháng nhau: Mạng tạo sinh (Generator) tạo ra dữ liệu giả lập từ vec-tơ nhiễu ngẫu nhiên, và Mạng phân biệt (Discriminator) thực hiện chức năng phân biệt giữa dữ liệu thật và dữ liệu giả. Tuy nhiên, GAN được thiết kế cho dữ liệu có tính liên tục và cấu trúc lưới như hình ảnh. Do đó, khi áp dụng cho dữ liệu bảng, chúng gặp nhiều thách thức lớn như không mô hình hóa được cấu trúc xác suất rời rạc và các mối quan hệ phi tuyến giữa các kiểu tính chất hỗn hợp; dữ liệu bảng thường có phân phối đa chế độ (multi-modal) và thưa, khiến GAN thông thường dễ rơi vào hiện tượng collapse; thiếu tính cục bộ không gian nên khó xác định các mối quan hệ ràng buộc bậc cao giữa các thuộc tính.

Để xử lý được dữ liệu dạng bảng, một biến thể của GAN là TGAN (Tabular Generative Adversarial Networks) đã được đề xuất. TGAN được tinh chỉnh và thiết kế chuyên biệt để giải quyết các đặc thù của dữ liệu dạng bảng bằng cách học cách phân phối xác suất đồng thời giữa các cột dữ liệu liên tục và phân loại. Tuy nhiên, các kiến trúc TGAN truyền thống vẫn gặp khó khăn trong việc xử lý các tập dữ liệu có sự mất cân bằng lớp (class imbalance) trầm trọng và chưa tối ưu hoàn toàn trong việc duy trì mối liên kết phức tạp giữa các đặc trưng phân loại bậc cao.

Nhằm tận dụng được ưu điểm và khắc phục được nhược điểm của TGAN, Saad M. Darwish và Noha A. El-Shoafy đã đề xuất mô hình CT-TabGAN (Conditional-TabNet GAN) [4]. Đây là mô hình tạo sinh được xây dựng dựa trên nền tảng kiến trúc của TGAN tích hợp TabNet nhằm giải quyết triệt để bài toán mất cân bằng lớp trên dữ liệu dạng bảng. TabNet là một kiến trúc học sâu được thiết kế chuyên biệt cho dữ liệu dạng bảng, tích hợp các cơ chế chú ý (attention mechanisms) cho phép mô hình lựa chọn những đặc trưng phù hợp nhất tại mỗi bước ra quyết định. Cải tiến của CT-TabGAN bao gồm hai điểm chính.

Thứ nhất, bộ tạo sinh tích hợp TabNet (TabNet-aware generator): CT-TabGAN tích hợp một lớp cấu trúc đặc trưng nhận biết TabNet để các đặc trưng tạo ra được sắp xếp tương thích với cơ chế của TabNet. Cụ thể, bộ tạo sinh bao gồm một mô-đun chuyển đổi đặc trưng đa nhánh mô phỏng các bộ chuyển đổi đặc trưng của TabNet để tạo ra các nhúng trung gian, phản ánh cách TabNet phân rã thông tin qua các bước

ra quyết định. Sau đó, lớp chiếu tạo độ thừa tái tổ hợp các nhúng trung gian này thành mẫu dữ liệu tổng hợp cuối cùng, đồng thời thúc đẩy việc nhóm các đặc trưng tương tự như cơ chế mật nạ của TabNet. Nhờ sự tương thích về cấu trúc này, dữ liệu được tạo ra không chỉ giữ được các tính chất thống kê cơ bản của dữ liệu gốc mà còn mô phỏng được cơ chế lựa chọn đặc trưng của TabNet.

Thứ hai, mô-đun lấy mẫu có điều kiện nâng cao của CT-TabGAN được thiết kế để bảo toàn tốt hơn các mối quan hệ phụ thuộc phức tạp giữa biến phân loại và biến liên tục, đặc biệt là trong dữ liệu dạng bảng có số chiều lớn. CT-TabGAN sử dụng bộ mã hóa điều kiện phân cấp (hierarchical conditional encoder) thay cho mã hóa one-hot. Bộ mã hóa này sẽ chuyển đổi từng giá trị phân loại thành một nhúng dày đặc (dense embedding) được học từ phân phối đồng thời (joint distribution) của tập dữ liệu. Các nhúng này sau đó được đưa qua một mạng học phụ thuộc, là nơi tạo ra các đặc trưng chéo bằng cơ chế chú ý (attention mechanisms) để tạo ra một biểu diễn điều kiện theo ngữ cảnh. Cuối cùng, vec-tơ điều kiện nâng cao này sẽ được đưa vào nhiều tầng khác nhau của bộ tạo sinh thay vì chỉ ở tầng đầu vào, đảm bảo rằng ngữ cảnh phân loại sẽ tác động xuyên suốt đến cả giai đoạn đầu và giai đoạn sau của quá trình tổng hợp đặc trưng.

2.2 Dữ liệu thực nghiệm

Bộ dữ liệu được sử dụng trong bài báo này là kết quả thi môn Toán của 395 HS và kết quả thi môn tiếng Bồ Đào Nha của 649 HS tại hai trường trung học Gabriel Pereira và Mousinho da Silveira tại Bồ Đào Nha, gồm 33 thuộc tính, với G3 là điểm thi cuối kỳ, như trong [1]. Điểm số G3 được mã hóa thành kết quả «hông» nếu bé hơn 10, và «đạt» nếu ngược lại. Như vậy, hai tập dữ liệu Math và Portuguese mất cân bằng, với tỷ lệ hỏng: đạt lần lượt là 130:265 và 100:549.

2.3 Thực nghiệm đánh giá hiệu suất dự đoán kết quả học tập của học sinh

Quá trình thực nghiệm để đánh giá hiệu suất dự đoán kết quả thi của HS được chúng tôi thực hiện như sau: Dữ liệu được thực hiện kiểm chứng chéo 5-fold (5-fold cross-validation). Ở mỗi lần lặp, mô hình CT-TabGAN sinh thêm phần tử cho tập huấn luyện, huấn luyện tập huấn luyện mới có tích hợp phần tử mới sinh và đánh giá. Trung bình cộng giá trị độ đo hiệu suất của kết quả phân lớp cho tập kiểm tra trong 5 lần lặp này chính là kết quả để đánh giá hiệu suất của mô hình. Ở đây, số lượng phần tử sinh thêm được chọn sao cho kết quả thu được là tốt nhất. Chúng tôi sử dụng Python 3.7 với các thư viện TensorFlow, PyTorch và Scikit-learn để cài đặt và thực nghiệm.

Ngoài ra, chúng tôi thực hiện thêm lựa chọn đặc

trung dựa vào Độ lợi thông tin (IG), chỉ giữ lại các thuộc tính có $IG > 0$ và so sánh kết quả trong trường hợp có và không có sử dụng CT-TabGAN để kiểm chứng hiệu quả của mô hình này.

Để đánh giá hiệu quả của quá trình phân lớp, trong bài báo này, các độ đo Precision (Độ chính xác), Recall (Độ hồi tưởng), F1 và AUC-ROC được sử dụng thêm cho độ chính xác toàn thể (Accuracy) như đối với việc phân lớp tập dữ liệu cân bằng. Khi phân lớp tập dữ liệu không cân bằng, chỉ sử dụng độ chính xác toàn thể không thích hợp để xác định tính hiệu quả của mô hình phân lớp vì các bộ phân lớp thông thường sẽ cho ra độ chính xác toàn thể rất cao do phần lớn các phần tử đều được gán nhãn lớp là lớp đa số và rất hiếm phần tử được gán nhãn lớp của lớp thiểu số.

2.4 Kết quả thực nghiệm

Thực nghiệm trên hai tập dữ liệu Math và Portuguese, bằng các phương pháp khác nhau: Baseline, CT-TabGAN theo các độ đo Precision, Recall và F1 cho kết quả như sau: Đối với môn Toán, các giá trị Precision, Recall, F1, Accuracy lần lượt của phương pháp Baseline là 73.39%, 60.77%, 65.75%, 79.75%; CT-TabGAN là 84.99%, 75.38%, 79.35%, 87.34%; Với môn tiếng Bồ Đào Nha, các giá trị Precision, Recall, F1, Accuracy lần lượt của phương pháp Baseline là 74.04%, 62.00%, 66.96%, 90.60%; CT-TabGAN là 73.67%, 72.00%, 72.41%, 91.68%.

Khi lựa chọn đặc trưng sử dụng IG, thu được 12 thuộc tính có IG bằng 0. Các thuộc tính school, sex, address, famsize, Pstatus, Mjob, Fjob, reason, guardian, failures, schoolsup, famsup, paid, activities, nursery, higher, internet, romantic, G1, G2 được giữ lại cho cả hai bộ dữ liệu Math và Portuguese. Thực hiện dự đoán trên hai tập dữ liệu sau khi lựa chọn đặc trưng bằng hai mô hình có và không có sử dụng CT-TabGAN, thu được kết quả: Đối với tập dữ liệu Math, phương pháp không sử dụng CT-TabGAN là FS-Baseline, các giá trị Precision, Recall, F1, Accuracy lần lượt là: 75.61%, 57.69%, 65.08%, 80.00%; phương pháp sử dụng CT-TabGAN là FS-CT-TabGAN, có kết quả đối với các chỉ số đánh giá lần lượt là: 77.83%, 69.23%, 73.10%, 83.29%. Đối với tập dữ liệu Portuguese, các giá trị Precision, Recall, F1, Accuracy cho hai phương pháp FS-Baseline và FS-CT-TabGAN tương ứng lần lượt 70.53%, 68.00%, 68.13%, 90.13%; và 72.30%, 72.00%, 71.38%, 91.36%.

Như vậy, kết quả khi áp dụng mô hình CT-TabGAN cho các tập dữ liệu đều có sự cải tiến. Đối với tập Math, giá trị Precision tăng 11.59%, Recall tăng 14.61%, F1 tăng 13.61%, Accuracy tăng 7.6% khi chưa áp dụng lựa chọn đặc trưng; trường hợp có áp dụng lựa chọn đặc trưng, Precision tăng 2.2%, Recall tăng 11.54%, F1 tăng 8%. Tương tự, đối với

tập dữ liệu Portuguese, khi chưa áp dụng lựa chọn đặc trưng, Recall tăng 10%, F1 tăng 5.4%, Accuracy tăng 1.07%; kết hợp với lựa chọn đặc trưng, kết quả có sự thay đổi như sau: Precision tăng 1.76%, Recall tăng 4%, F1 tăng 3.24%, Accuracy tăng 1.23%. Điều này cho thấy khả năng không bỏ sót các trường hợp học sinh cần được tư vấn đối với môn Toán và tiếng Bồ Đào Nha có sự cải thiện đáng kể so với việc không sử dụng CT-TabGAN.

Bên cạnh đó, giá trị AUC-ROC khi áp dụng CT-TabGAN cho tập dữ liệu Math tăng từ 0.847 lên 0.948 trường hợp không áp dụng lựa chọn đặc trưng và từ 0.886 lên 0.925 trường hợp có lựa chọn đặc trưng. Đối với tập dữ liệu Portuguese, các độ tăng này tương ứng là từ 0.927 lên 0.937 và 0.925 lên 0.933. Các giá trị AUC-ROC sau khi áp dụng CT-TabGAN đều lớn hơn 0.9, cho thấy mô hình có hiệu suất xuất sắc. Kết quả này một lần nữa khẳng định khả năng dự đoán hiệu quả của mô hình.

3. Kết luận

Trong bài báo này, chúng tôi đã trình bày phương pháp dự đoán kết quả thi cuối kỳ cho HS, sử dụng mô hình CT-TabGAN để sinh thêm phần tử, xử lý vấn đề mất cân bằng dữ liệu. Thực nghiệm trên hai tập dữ liệu mẫu thường được sử dụng để dự đoán kết quả môn Toán và môn tiếng Bồ Đào Nha được thực hiện và so sánh với Tabnet cơ bản, trường hợp có và không có lựa chọn đặc trưng. Kết quả thực nghiệm đã cho thấy mô hình CT-TabGAN có hiệu quả tốt dựa trên các giá trị độ đo đánh giá hiệu suất, làm tăng khả năng dự đoán kết quả thi hòng cho HS đối với hai môn Toán và tiếng Bồ Đào Nha. Tuy nhiên, kết quả này vẫn còn hạn chế do kích thước của tập dữ liệu dùng cho thực nghiệm chưa đầy đủ cho tất cả các môn học, số lượng mẫu cũng như tỉ lệ mất cân bằng giữa các lớp vẫn chưa lớn. Các nghiên cứu trong tương lai có thể phát triển theo hướng áp dụng cho các tập dữ liệu có số lượng mẫu lớn hơn, tỉ lệ mất cân bằng cao hơn và cho các môn học khác để tăng tính toàn diện cho nghiên cứu.

Tài liệu tham khảo

- [1] P. Cortez and A. Silva, "Using data mining to predict secondary school student performance," in Proceedings of 5th Future Business Technology Conference (EUROSIS), 2008, pp. 5–12.
- [2] A. Hashim, W. Akeel, and A. Khalaf, "Student Performance Prediction Model based on Supervised Machine Learning Algorithms," IOP Conference Series: Materials Science and Engineering, vol. 928, no. 3, p. 032018, 2020.
- [3] H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.