Journal of educational equipment: Education management, Volume 1, Issue 314(June 2024)
ISSN 1859 - 0810

Một phương pháp cải thiện hiệu suất dự đoán kết quả thi của học sinh bằng kỹ thuật học máy

*Nguyễn Thị Lan Anh, Nguyễn Lê Trung Thành, Vĩnh Anh Nghiêm Quân**
**Trường Đại học sư phạm, Đại học Huế*
Received: 13/03/2024; Accepted: 26/03/2024; Published: 16/04/2024

Abstract: *Predicting student performance to support and counsel at the right time is one of the most important problems that many researchers care about. Though many machine learning based methods have been proposed to identify at-risk students, the prediction results are still limited. In this paper, we suggest a method based on feature selection and over-sampling to improve the performance. The experimental results show that our approach is comparable to other methods.*

Keywords: *Student performance prediction, machine learning, over-sampling, feature selection*

1. Đặt vấn đề

Giáo dục không chỉ là chìa khóa thành công cho mỗi cá nhân trong xã hội mà còn là nền tảng để xã hội phát triển và thịnh vượng. Bên cạnh đó, đối tượng học sinh lại là rường cột của nước nhà, là chủ nhân đất nước trong tương lai. Cho nên, làm sao để học sinh có thể nhận được một nền tảng giáo dục phù hợp nhằm phát huy tối đa năng lực của bản thân là một trong những vấn đề được các nhà giáo dục và hoạch định chính sách quan tâm.

Một trong số các vấn đề liên quan đến việc phát huy năng lực cá nhân của học sinh là làm sao phát hiện sớm khả năng thi hỏng của học sinh để kịp thời tư vấn và hỗ trợ cho các em, vì điểm thi cuối kỳ của mỗi môn học có ảnh hưởng rất lớn đến kết quả toàn thể của mỗi học sinh trong năm học hiện tại và do đó có khả năng ảnh hưởng đến kết quả trong các giai đoạn tiếp theo của các em.

Với sự phát triển của công nghệ, đặc biệt là sự phát triển của lĩnh vực học máy, các chương trình tư vấn, dự đoán đã ra đời và được ứng dụng để giải quyết nhiều bài toán thực tế. Nhiều nghiên cứu về ứng dụng học máy để dự đoán kết quả học tập của học sinh cũng như để phân tích các yếu tố ảnh hưởng đến kết quả này đã được thực hiện [1][2] [3]. Tuy nhiên, kết quả đạt được vẫn còn hạn chế. Một trong những lý do dẫn đến sự hạn chế này là sự mất cân đối dữ liệu, hay còn gọi là mất cân bằng dữ liệu. Dữ liệu mất cân bằng gây ra khó khăn khi phân lớp bằng các thuật toán phân lớp truyền thống, vì 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 [4]. Hay nói cách khác, hầu hết các phần tử thuộc lớp đa số (lớp có số lượng phần tử lớn) sẽ được phân lớp đúng và các phần tử thuộc lớp thiểu số (lớp có số lượng phần tử ít hơn) cũng sẽ được gán nhãn lớp là nhãn lớp của lớp đa số, kết quả là accuracy của việc phân lớp rất cao trong khi sensitivity lại rất thấp. Khi áp dụng vào bài toán dự đoán kết quả thi của học sinh, những học sinh có khả năng thi hỏng lại được dự đoán là đạt, nên không được tư vấn kịp thời để cải thiện kết quả, dẫn đến hậu quả đáng tiếc.

Các nghiên cứu về dữ liệu mất cân bằng trong những năm gần đây nhận được rất nhiều quan tâm và được ứng dụng trong khá nhiều lĩnh vực khác nhau như phát hiện gian lận tài chính, dự đoán cấu trúc protein, dự đoán tương tác giữa protein-protein, phân lớp microRNA, chẩn đoán bệnh trong y học…, nhằm mục đích phát hiện các đối tượng hiếm nhưng quan trọng. Nhiều phương pháp nâng cao hiệu quả bài toán phân lớp dữ liệu mất cân bằng đã được đề xuất, 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; 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; và nhóm các phương pháp kết hợp [5].

Theo một số tác giả, 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 [5]. Do đó, trong bài báo này, chúng tôi áp dụng một phương pháp thuộc nhóm các phương pháp sinh thêm phần tử cho lớp thiểu số để nâng cao hiệu quả dự đoán kết quả thi cho học sinh.

2. Nội dung nghiên cứu

2.1 Phương pháp nâng cao hiệu suất bài toán phân lớp dữ liệu mất cân bằng

2.1.1 Lựa chọn đặc trưng

Lựa chọn đặc trưng hay lựa chọn thuộc tính, trong những nghiên cứu gần đây được cho thấy là có tác dụng trong việc cải thiện hiệu suất bài toán phân lớp dữ liệu mất cân bằng [6]. Ý tưởng cơ bản của kỹ thuật này là chỉ giữ lại một tập con các thuộc tính quan trọng của dữ liệu thay vì sử dụng tất cả các đặc trưng để thực hiện phân lớp. Có nhiều kỹ thuật lựa chọn đặc trưng khác nhau như Filter, Wrapper hay phương pháp lai. Một trong những phương pháp lựa chọn đặc trưng phổ biến thuộc nhóm Filter là sử dụng Độ lợi thông tin (Information Gain) để tính trọng số của các thuộc tính và giữ lại các thuộc tính có giá trị nhất dựa vào giá trị ngưỡng cho trước.

Với tập dữ liệu S, tập các thuộc tính A, Độ lợi thông tin $IG(A_m)$ của thuộc tính $A_m \in A$ được xác định như sau:

$$IG(A_m) = H(S)-H(S|A_m) \quad (1)$$

Trong đó:

H(S) là entropy của tập dữ liệu S

$H(S|A_m)$ là entropy có điều kiện của S với thuộc tính A_m

2.1.2 Thuật toán sinh thêm phần tử cho lớp thiểu số SMOTE-NC

Một trong những phương pháp Sinh thêm phần tử phổ biến cho tập dữ liệu mất cân bằng $S = S_{min} \cup S_{max}$ (S_{min} là lớp thiểu số, S_{max} là lớp đa số), có tập thuộc tính A chứa các đặc trưng dạng số và định danh, được áp dụng trong nhiều nghiên cứu khác nhau là phương pháp SMOTE-NC [7]. Phương pháp này cho phép các phần tử lớp thiểu số sinh thêm phần tử mới nằm giữa nó và một trong những phần tử láng giềng cùng nhãn lớp gần nhất của nó, được mô tả như bên dưới:

Với mỗi phần tử $x_i \in S_{min}$, tìm k láng giềng gần nhất của x_i trong S_{min}, nghĩa là k phần tử trong S_{min} có khoảng cách đến x_i là bé nhất. Khoảng cách ở đây là khoảng cách Euclid, được xác định dựa vào các thuộc tính liên tục, đối với mỗi thuộc tính định danh thì giá trị trung vị độ lệch chuẩn của tất cả các thuộc tính liên tục của lớp thiểu số được sử dụng.

Chọn ngẫu nhiên một phần tử x_n trong số k láng giềng tìm được.

Phần tử mới được sinh ra x_{new} xác định bởi công thức:

Với những thuộc tính liên tục A_m, m =1..|A|:

, δ là một giá trị ngẫu nhiên thuộc [0,1]

Với các thuộc tính định danh, nhận giá trị là giá trị xuất hiện nhiều nhất trên thuộc tính A_m trong k láng giềng của phần tử x_i đang xét.

x_{new} sẽ được gán nhãn lớp là nhãn lớp của lớp thiểu số.

2.2 Dữ liệu thực nghiệm

Trong bài báo này, chúng tôi sử dụng bộ dữ liệu UCI liên quan đến kết quả thi môn Toán của 395 học sinh và kết quả thi môn tiếng Bồ Đào Nha của 649 học sinh tại hai trường trung học Gabriel Pereira và Mousinho da Silveira tại Bồ Đào Nha, với 33 thuộc tính như trong [1]. Học sinh có điểm cuối kỳ G3 bé hơn 10 thì hỏng, ngược lại là đạt. Mục đích của chúng tôi trong bài báo này là dự đoán để phát hiện các học sinh có khả năng thi hỏng hay không. Như vậy, tỷ lệ hỏng: đạt của tập dữ liệu Math là 130:265 và của tập Portuguese là 100:549, và cả hai tập dữ liệu này đều mất cân bằng.

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 học sinh được chúng tôi thực hiện như sau:

1. Độ lợi thông tin (IG) của các thuộc tính được xác định bằng hàm information_gain của gói FselectorRcpp trong R. Sau đó, các thuộc tính có giá trị IG lớn hơn 0 được giữ lại. Quá trình này giúp làm giảm sự ảnh hưởng của các thuộc tính không quan trọng đến kết quả dự đoán.

2. Dữ liệu sau xử lý được thực hiện kiểm chứng chéo 5-fold (5-fold crossvalidation). Ở mỗi lần lặp, lớp thiểu số của tập huấn luyện tương ứng được sinh thêm phần tử để đảm bảo có số lượng tương đương với số lượng phần tử của lớp đa số bằng hàm smotenc, k=20, của gói lệnh Themis trong R. Lớp thiểu số mới sau đó được kết hợp với lớp đa số để tạo ra tập huấn luyện mới. 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, máy vector hỗ trợ (KSVM) thuộc gói lệnh kernlab được sử dụng làm bộ phân lớp chính.

Để đánh giá hiệu quả của quá trình phân lớp, trong bài báo này, các độ đo Sensitivity (Độ nhạy), Specificity (Độ đặc hiệu) và G-mean được sử dụng thay 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, Độ chính xác toàn thể này 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. Các giá trị Sensitivity, Specificity, G-mean được định nghĩa như sau:

(2) , (3)

(4)

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

2.4 Kết quả thực nghiệm

Sử dụng hàm information_gain của gói FselectorRcpp trong R thu được độ lợi thông tin (IG) của các thuộc tính. Thuộc tính G2 có IG cao nhất cho tập dữ liệu Math và Portuguese lần lượt là 0.44241 và 0.26123. Hai thuộc tính khác có IG cao nhì và ba là G2 và số lần thi hỏng. Ba thuộc tính có IG cao nhất này có ảnh hưởng lớn nhất đến kết quả thi cuối cùng G3, đối với cả hai tập dữ liệu. Các yếu tố như trình độ học vấn của cha mẹ, thời gian di chuyển từ nhà đến trường, thời gian học hàng tuần, chất lượng các mối quan hệ gia đình, thời gian rãnh sau giờ học, tần suất ra ngoài với bạn, việc tiêu thụ bia rượu trong ngày bình thường hoặc cuối tuần, tình trạng sức khỏe hiện tại, số lần vắng học có IG bằng 0, tức là không có ảnh hưởng gì đến việc thi hỏng hay đạt của học sinh trong trường hợp này.

Chúng tôi tính kết quả dự đoán kết quả thi của học sinh cho hai tập dữ liệu Math và Portuguese, bằng các phương pháp khác nhau, trước khi áp dụng lựa chọn đặc trưng KSVM, SMOTENC và sau khi lựa chọn đặc trưng: FS-KSVM, FS-SMOTENC theo các độ đo Sensitivity, Specificity và G-mean. Đối với môn Toán, các giá trị Sensitivity, Specificity, G-mean, Acuracy lần lượt của phương pháp KSVM là 83.85, 93.84, 88.70, 90.58; phương pháp SMOTENC là 91.92, 88.99, 90.43, 89.95; phương pháp FS-KSVM là: 90.15, 94.03, 92.07, 92.76; của phương pháp FS-SMOTENC là: 96.92, 90.15, 93.45, 92.36. Với môn tiếng Bồ Đào Nha, các giá trị Sensitivity, Specificity, G-mean, Acuracy lần lượt của phương pháp KSVM là 62, 97.45, 77.73, 91.97; phương pháp SMOTENC là 80.4, 93.50, 86.7, 91.48; phương pháp FS-KSVM là: 73.6, 97.23, 84.57, 93.58; của phương pháp FS-SMOTENC là: 85.6, 93.8, 89.56, 92.53.

So sánh kết quả khi áp dụng phương pháp sinh thêm phần tử FS-SMOTENC kết hợp với lựa chọn đặc trưng với kết quả của các phương pháp khác cho thấy hiệu quả của nó trong xử lý sự mất cân bằng dữ liệu. Cụ thể, đối với tập dữ liệu Math, Sensitivity và G-mean của FS-SMOTENC đã tăng lần lượt 13.077% và 4.748% so với KSVM, tăng 5% và 3.012% so với SMOTENC, tăng 6.769% và 1.381% so với FS-KSVM. Đối với tập dữ liệu Portuguese, Sensitivity của FS-SMOTENC tăng lần lượt 23.6%, 5.2%,12% so với các phương pháp KSVM, SMOTENC, FS-KSVM và Gmean của FS-SMOTENC lần lượt tăng 11.83%, 2.854%, 4.984% so với KSVM, SMOTENC, FS-KSVM.

Đối với độ chính xác toàn cục, so với phương pháp của Cortez và cộng sự [1], cùng sử dụng Máy vector hỗ trợ làm bộ phân lớp chính, phương pháp của chúng tôi đề xuất tăng 6.062% trên tập dữ liệu Math và 1.131% trên tập dữ liệu Portuguese.

3. Kết luận

Trong bài báo này, chúng tôi đã trình bày một phương pháp làm tăng hiệu quả dự đoán kết quả thi cuối kỳ cho học sinh, dựa vào kỹ thuật lựa chọn đặc trưng và sinh thêm phần tử để xử lý vấn đề mất cân bằng dữ liệu. Chúng tôi cũng đã thực hiện đánh giá hiệu suất của phương pháp này trên hai tập dữ liệu mẫu thường được sử dụng để dự đoán kết quả môn Toán và môn tiếng Bồ Đào Nha và so sánh với các thuật toán khác. Kết quả thực nghiệm đã cho thấy thuật toán được đề xuất có hiệu quả tốt dựa trên các giá trị độ đo đánh giá hiệu suất Sensitivity và G-mean. Dù vậy, do kích thước của tập dữ liệu dùng để thực nghiệm vẫn còn hạn chế, trong tương lai, chúng tôi sẽ tìm kiếm các tập dữ liệu thực khác để thực hiện việc kiểm thử đầy đủ hơn, từ đó có thể cho ra những kết quả chính xác và tin cậy hơn. Phương pháp này cũng có thể áp dụng để làm tăng hiệu quả dự đoán kết quả các môn học khác.

Tài liệu tham khảo

1. Cortez, Paulo & Silva, Alice, *"Using data mining to predict secondary school student performance,"* *EUROSIS,* 2008.

2. Ying, Dahao & Ma, Jieming, *"Student Performance Prediction with Regression Approach and Data Generation," Applied Sciences,* vol. 14, p. 1148, 2024.

3. Hashim, Ali & Akeel, Wid & Khalaf, Alaa, "Student Performance Prediction Model based on Supervised Machine Learning Algorithms," *IOP Conference Series: Materials Science and Engineering*, 2020.

4. Hashim, Ali & Akeel, Wid & Khalaf, Alaa, "Student Performance Prediction Model based on Supervised Machine Learning Algorithms," *IOP Conference Series: Materials Science and Engineering*, 2020.