Tạp chí Khoa học Trường Đại học Cần Thơ
Phần A: Khoa học Tự nhiên, Công nghệ và Môi trường
website: sj.ctu.edu.vn

DOI:10.22144/ctu.jvn.2020.049

DỰ BÁO KẾT QUẢ HỌC TẬP BẰNG KỸ THUẬT HỌC SÂU VỚI MẠNG NƠ-RON ĐA TẦNG

Lưu Hoài Sang[1*], Trần Thanh Điện[2], Nguyễn Thanh Hải[3] và Nguyễn Thái Nghe[3]

[1]*Trung tâm Công nghệ Phần mềm, Trường Đại học Cần Thơ*

[2]*Nhà xuất bản Đại học Cần Thơ, Trường Đại học Cần Thơ*

[3]*Khoa Công nghệ Thông tin và Truyền thông, Trường Đại học Cần Thơ*

**Người chịu trách nhiệm về bài viết: Lưu Hoài Sang (email: lhsang@ctu.edu.vn)*

Thông tin chung:
Ngày nhận bài: 26/05/2020
Ngày nhận bài sửa: 19/06/2020
Ngày duyệt đăng: 29/06/2020

Title:
Predicting student's performance through deep learning using a multi-layer perceptron

Từ khóa:
Học máy, học sâu, khai thác dữ liệu, mạng nơ-ron truyền thẳng đa tầng

Keywords:
Data mining, deep learning, machine learning, student's performance

ABSTRACT

Student performance prediction is one of the most concerned issues in the field of education and training. The prediction of courses results enables students to select courses appropriately. Moreover, this helps education managers and lecturers to indicate the students who should be monitored and supported to complete the courses with good results. Therefore, the student performance prediction is expected to reduce formal warnings and expulsions from universities due to students' poor performance. In this study, a method was proposed to predict student performance using deep learning techniques exploring and mining data from universities' student information system. From collected data, the data was analyzed and pre-processed before fetching them into a multi-layer perceptron to do prediction tasks. The obtained results from the proposed model reveal rather accurate forecasts; hence, the model is expected to apply in practical cases.

TÓM TẮT

Dự báo kết quả học tập là một chủ đề đang được quan tâm trong lĩnh vực giáo dục đào tạo. Dự báo sớm kết quả học tập có thể giúp sinh viên lựa chọn học phần phù hợp với năng lực cá nhân, giúp nhà quản lý và giảng viên xác định được những sinh viên cần được quan tâm hỗ trợ nhiều hơn để hoàn thành tốt học phần, giảm tình trạng cảnh báo học vụ hoặc buộc thôi học do kết quả học tập kém, từ đó tiết kiệm được thời gian chi phí cho cả sinh viên, gia đình, nhà trường và xã hội. Bài viết này đề xuất một phương pháp dự báo kết quả học tập của sinh viên bằng kỹ thuật học sâu nhằm khai thác cơ sở dữ liệu trong hệ thống quản lý sinh viên tại các trường đại học. Dữ liệu sau khi thu thập được phân tích, tiền xử lý dữ liệu, thiết kế và huấn luyện mạng nơ-ron đa tầng. Kết quả thực nghiệm cho thấy mô hình đề xuất cho kết quả dự đoán khá chính xác và hoàn toàn khả thi để áp dụng vào thực tế.

Trích dẫn: Lưu Hoài Sang, Trần Thanh Điện, Nguyễn Thanh Hải và Nguyễn Thái Nghe, 2020. Dự báo kết quả học tập bằng kỹ thuật học sâu với mạng nơ-ron đa tầng. Tạp chí Khoa học Trường Đại học Cần Thơ. 56(3A): 20-28.

1 GIỚI THIỆU

Thời gian gần đây, tình trạng sinh viên ở các viện, trường bị cảnh báo học vụ hoặc buộc thôi học đang có chiều hướng gia tăng. Chẳng hạn, tại Trường Đại học Cần Thơ, nếu như học kỳ 1 năm học 2018-2019, số sinh viên bị cảnh báo học vụ một học kỳ là 886 và hai học kỳ là 125 thì con số này trong học kỳ 1 năm học 2019-2020 lần lượt là 986 và 196 (Đại học Cần Thơ, 2020). Một trong những nguyên nhân chính dẫn đến kết quả học tập không tốt của sinh viên là do chưa lựa chọn đúng những học phần phù hợp với khả năng của mình. Điều này dẫn đến việc sinh viên phải học kéo dài thời gian học tập so với kế hoạch ban đầu, lãng phí thời gian, tiền của không chỉ của gia đình mà cả nhà trường và xã hội. Vì vậy, dự báo kết quả học tập của sinh viên là một chủ đề nghiên cứu quan trọng trong khai thác dữ liệu giáo dục được nhiều nhà nghiên cứu quan tâm (Guo *et al.*, 2015, Tanuar *et al.*, 2018, Altabrawee *et al.*, 2019).

Theo Thai-Nghe *et al.* (2011), dự báo kết quả học tập của sinh viên là công việc quan trọng trong khai thác dữ liệu giáo dục; kiến thức của sinh viên có thể được cải thiện và tích lũy theo thời gian. Từ ý tưởng này, một cách tiếp cận sử dụng kỹ thuật phân rã ma trận có ảnh hưởng bởi yếu tố thời gian (tensor factorization - TF) đã đề xuất để dự báo kết quả học tập của sinh viên. Với phương pháp này, nhóm tác giả có thể cá nhân hóa dự báo cho từng sinh viên cụ thể. Kết quả thực nghiệm trên hai tập dữ liệu lớn cho thấy việc kết hợp các kỹ thuật dự báo vào quá trình phân rã ma trận là một cách tiếp cận hiệu quả và đầy hứa hẹn.

Việc sử dụng thư viện mã nguồn mở phục vụ cho công việc dự báo cũng được sử dụng nhiều trong thời gian gần đây. Huynh-Ly and Thai-Nghe (2013) đã xây dựng hệ thống dự báo kết quả học tập của sinh viên sử dụng thư viện hệ thống gợi ý mã nguồn mở MyMediaLite. Với cơ sở dữ liệu điểm thu thập được từ hệ thống quản lý kết quả học tập, nhóm tác giả đã đề xuất sử dụng kỹ thuật kỹ thuật phân rã ma trận thiên vị (biased matrix factorization-BMF) để dự đoán kết quả học tập của sinh viên, từ đó làm cơ sở giúp sinh viên lựa chọn học phần phù hợp hơn.

Khả năng kết hợp giữa các phương pháp dự báo cũng được các nhà nghiên cứu tận dụng. Hai *et al.* (2015) đã xây dựng mô hình dự báo kết quả học tập của học sinh dựa trên sự kết hợp phương pháp gần đúng Taylor với các mô hình xám (grey models) để có thể đạt được các giá trị dự báo tối ưu nhất bằng cách tính gần đúng nhiều lần nhằm cải thiện độ chính xác của dự báo. Kết quả nghiên cứu giúp cho giáo viên và nhà quản lý giáo dục có giải pháp phù hợp nhằm cải thiện kết quả học tập của sinh viên có quá trình học tập không ổn định. Iqbal *et al.* (2017) sử dụng các kỹ thuật lọc cộng tác (collaborative filtering - CF), phân rã ma trận (matrix factorization - MF) và kỹ thuật restricted boltzmann machines (RBM) để phân tích một cách có hệ thống dữ liệu được thu thập từ một trường đại học. Kết quả cho thấy, kỹ thuật RBM dự báo kết quả học tập của sinh viên tốt hơn so với các kỹ thuật còn lại.

Thực tế, các giải thuật lọc cộng tác được sử dụng phổ biến trong các hệ thống gợi ý do tính đơn giản và hiệu quả. Tuy nhiên, độ thưa thớt của dữ liệu làm hạn chế tính hiệu quả của các giải thuật này và rất khó để cải thiện hơn nữa kết quả gợi ý. Do đó, các mô hình kết hợp các thuật toán gợi ý lọc cộng tác với công nghệ học sâu được quan tâm nhiều hơn. Zhang *et al.* (2018) đề xuất mô hình dựa trên mô hình hồi quy đa thức bậc 2 (quadratic polynomial regression model) thu được các đặc tả tiềm ẩn (latent features) chính xác hơn bằng cách cải thiện giải thuật phân rã ma trận truyền thống. Sau đó, các đặc tả tiềm ẩn làm dữ liệu đầu vào (input) của mô hình học sâu (deep neural network model). Thực nghiệm trên 3 tập dữ liệu cho thấy mô hình đề xuất cải thiện hiệu quả gợi ý rất tốt so với các mô hình gợi ý truyền thống.

Một số tiếp cận khác kết hợp mô hình lọc cộng tác với học sâu cũng được Fu *et al.* (2019) đề xuất. Với cách tiếp cận này, trong giai đoạn dự báo, một mạng nơ-ron truyền thẳng (feed-forward neural networks) dùng để mô phỏng sự tương tác giữa user và item, trong đó các véc-tơ biểu diễn ở giai đoạn tiền xử lý được sử dụng làm đầu vào của mạng thần kinh. Các thực nghiệm dựa trên hai bộ dữ liệu MovieLens 1M và MovieLens 10M được thực hiện để kiểm chứng tính hiệu quả của phương pháp này và cho kết quả rất khả quan.

Bài viết này đề xuất sử dụng kỹ thuật học sâu với mạng nơ-ron truyền thẳng đa tầng (multi layer perceptron - MLP) để xây dựng mô hình dự đoán kết quả học tập của sinh viên đối với các học phần mới dựa trên kết quả của các học phần trước đó đã học. Ngoài ra, để cải thiện kết quả dự đoán, chúng tôi cũng xem xét một số thông tin khác như điểm tuyển sinh, ngành học, giảng viên,… để đưa vào mô hình đề xuất.

2 MẠNG NƠ-RON ĐA TẦNG

Mô hình mạng nơ-ron thường được sử dụng rộng rãi, nhất là mô hình MLP. Một mạng MLP tổng quát

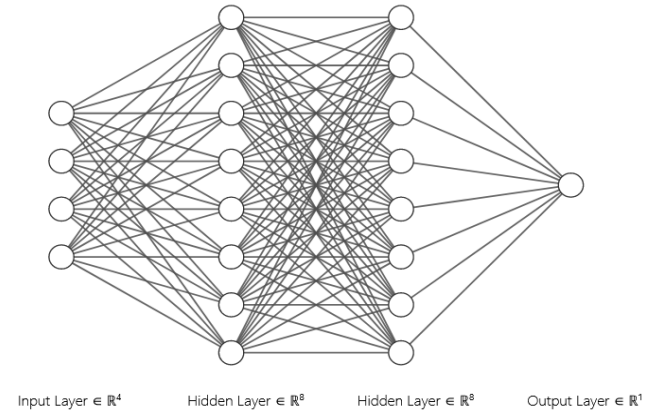
là mạng có n tầng ($n \geq 2$), trong đó bao gồm một tầng đầu vào, một tầng đầu ra và một hoặc nhiều tầng ẩn được minh họa như Hình 1.

Kiến trúc của một mạng MLP tổng quát có thể mô tả như sau:

– Đầu vào là các vector (x_1, x_2, ..., x_p) trong không gian p chiều, đầu ra là các vector (y_1, y_2, ..., y_q) trong không gian q chiều. Đối với các bài toán phân loại, p chính là kích thước của mẫu đầu vào, q chính là số lớp cần phân loại. Xét ví dụ trong bài toán nhận dạng chữ số: với mỗi mẫu ta lưu tọa độ (x,y) của 8 điểm trên chữ số đó, nhiệm vụ của mạng là phân loại các mẫu này vào một trong 10 lớp tương ứng với 10 chữ số 0, 1, …, 9. Khi đó p là kích thước mẫu và bằng 8 x 2 = 16; q là số lớp và bằng 10.

– Mỗi nơ-ron thuộc tầng sau liên kết với tất cả các nơ-ron thuộc tầng liền trước nó. Đầu ra của nơ-ron tầng trước là đầu vào của nơ-ron thuộc tầng liền sau nó.

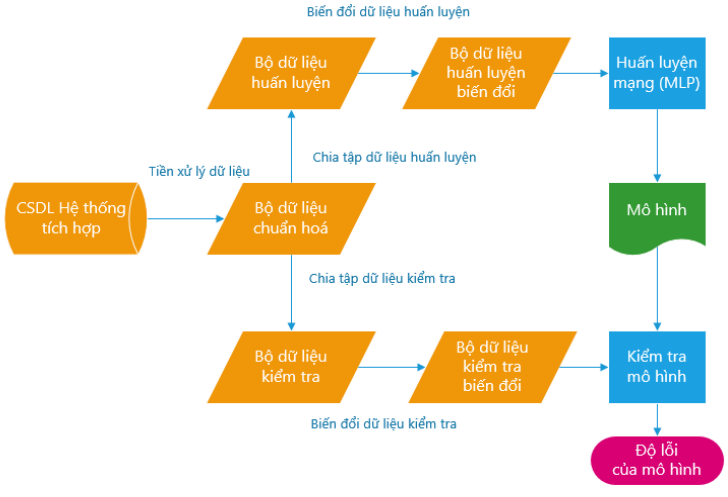
– Mạng MLP hoạt động như sau: tại tầng vào, các nơ-ron nhận tín hiệu vào xử lý (tính tổng trọng số, gửi tới hàm kích hoạt) rồi cho ra kết quả (là kết quả của hàm kích hoạt); kết quả này sẽ được truyền tới các nơ-ron thuộc tầng ẩn thứ nhất; các nơ-ron tại đây tiếp nhận như là tín hiệu đầu vào, xử lý và gửi kết quả đến tầng ẩn thứ 2;…; quá trình tiếp tục cho đến khi các nơ-ron thuộc tầng ra cho kết quả.



Hình 1: Minh họa một mạng MLP gồm 2 tầng ẩn và 2 tầng đầu vào và đầu ra

3 XÂY DỰNG MÔ HÌNH DỰ BÁO KẾT QUẢ HỌC TẬP

Lược đồ chung cho giải pháp đề xuất được trình bày trong Hình 2. Từ việc thu thập được của Hệ thống quản lý sinh viên, dữ liệu tiến hành tiền xử lý, huấn luyện mạng MLP và đánh giá kết quả. Các bước thực hiện sẽ được trình bày chi tiết trong phần tiếp theo.



Hình 2: Lược đồ chung của giải pháp đề xuất

3.1 Thu thập và tiền xử lý dữ liệu

Để kiểm chứng mô hình đề xuất, nghiên cứu sử dụng dữ liệu thu thập từ kết quả điểm môn học của sinh viên hệ đào tạo chính quy của Trường Đại học Cần Thơ, tuy nhiên mô hình này hoàn toàn có thể sử dụng cho các đơn vị đào tạo khác. Dữ liệu thu thập bắt đầu từ năm học 2007-2008 đến năm học 2018-2019 với số lượng mẫu được khảo sát là 3.828.879 trên kết quả 4.699 môn học của 83.993 sinh viên thuộc 16 đơn vị với số lượng chi tiết như Bảng 1.

Bảng 1: Dữ liệu khảo sát phân bổ theo khoa

Stt	Tên khoa	Số lượng mẫu	Số Sinh viên	Số Môn học
1	Kinh tế	689.930	15.997	3.754
2	Công nghệ	633.545	14.138	1.877
3	Nông nghiệp	473.736	9.744	2.094
4	Sư phạm	371.284	6.867	2.006
5	Công nghệ Thông tin và Truyền thông	219.808	6.000	1.509
6	Phát triển Nông thôn	204.033	4.943	756
7	Luật	209.293	3.986	2.905
8	Môi trường và Tài nguyên thiên nhiên	209.346	4.244	1.101
9	Ngoại ngữ	182.441	3.860	2.654
10	Thủy sản	159.135	3.776	899
11	Khoa học Xã hội và Nhân văn	143.960	3.299	1.079
12	Khoa học Tự nhiên	121.489	2.736	884
13	Viện NC&PT Công nghệ Sinh học	76.631	1.524	797
14	Khoa học Chính trị	63.502	1.381	306
15	Viện Nghiên cứu phát triển ĐBSCL	38.001	806	231
16	Bộ môn Giáo dục thể chất	32.745	692	282
	Tổng cộng	3.828.879	83.993	

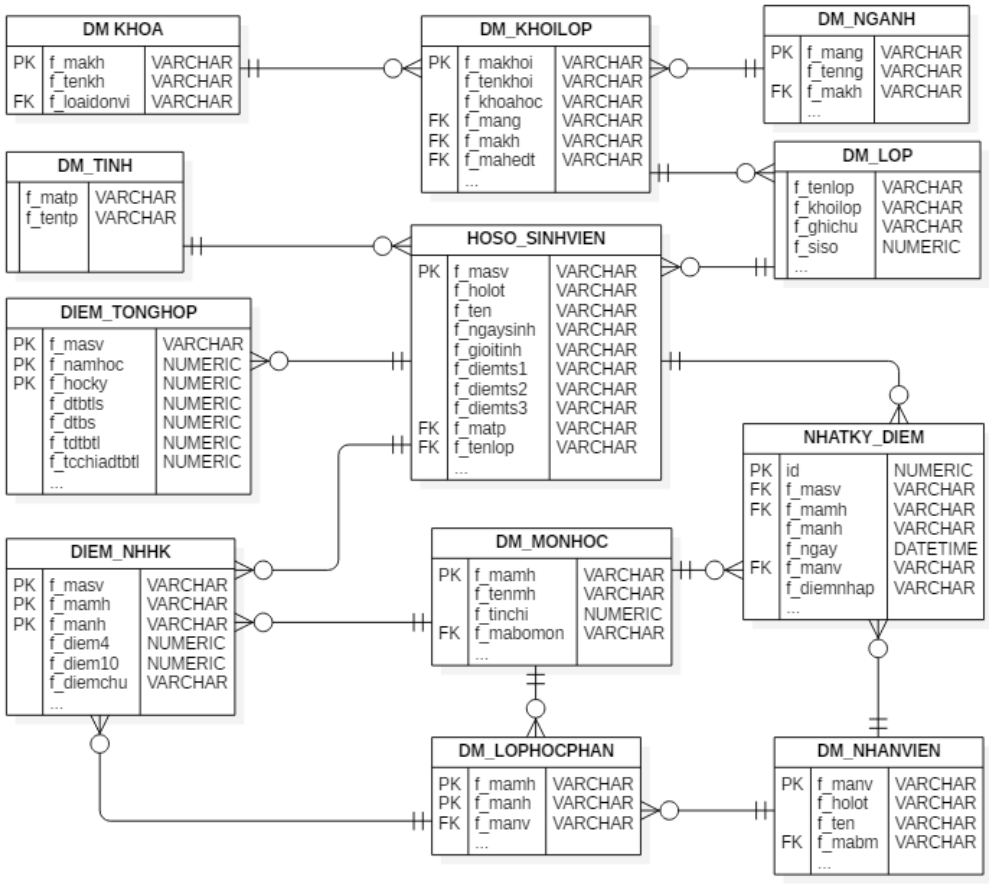
Trong thực tế, các hệ thống quản lý kết quả học tập của sinh viên có rất nhiều thuộc tính như trình bày một phần trong Hình 3. Dựa trên các nghiên cứu trước đây (Nguyen Thai *et al.*, 2007) và tiền thực nghiệm, nghiên cứu này chỉ giữ lại một số thuộc tính (điểm trung bình, ngành học, giới tính, điểm tuyển sinh đầu vào, hộ khẩu) và bổ sung một số thuộc tính quan trọng ảnh hưởng đến dự đoán kết quả học tập, chi tiết các bước tiền xử lý được mô tả dưới đây.

– Bước 1: Loại bỏ các thuộc tính dữ liệu không liên quan như học tên sinh viên, tên môn học, tên giảng viên, lịch học, ngày vào đoàn/đảng…

– Bước 2: Thực hiện loại bỏ dữ liệu nhiễu và tiền xử lý đối với phần điểm của sinh viên: trong đó có các dữ liệu như điểm miễn (-2), điểm chưa hoàn tất học phần (-1), điểm rút học phần (-5), những trường hợp sinh viên đăng ký nhưng không tham gia học tập (*null*), những sinh viên không có điểm tuyển sinh đầu vào (18% dữ liệu nhiễu trên tập dữ liệu khảo sát). Trong trường hợp dữ liệu môn học là ở học kỳ đầu thì gán giá trị 0 cho điểm trung bình học kỳ, điểm trung bình tích lũy, tổng điểm tích lũy, tổng số tín chỉ tích lũy của học kỳ trước.

– Bước 3: Xử lý những thuộc tính không đủ thông tin: Ví dụ, đa số các học phần thực hành các khoa tự phân công người giảng dạy và không nhập để lưu trữ trên phần mềm nên những trường hợp không có người giảng dạy cho môn đó thì ID của người dạy sẽ gán là 0.

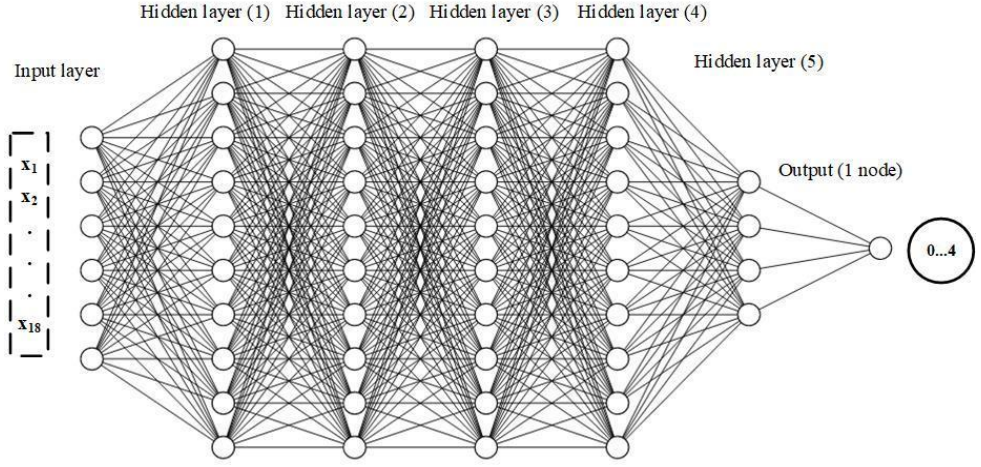
– Bước 4: Xử lý dữ liệu dạng chuỗi: Dữ liệu đa phần là kiểu chuỗi nên chúng tôi thực hiện "số hóa" dữ liệu. Ví dụ mã sinh viên được nhập ban đầu theo kiểu chuỗi, chúng tôi chuyển sang thành ID kiểu số.



Hình 3: Mô hình quan hệ thực thể (entity relationship diagram - ERD) trích ra từ hệ thống quản lý kết quả học tập

3.2 Xây dựng mô hình dự đoán dùng mạng MLP

Sau khi tiền xử lý và lựa chọn các thuộc tính phù hợp, mạng MLP được xây dựng như mô tả trong Hình 4.



Hình 4: Kiến trúc mạng MLP được đề xuất trong dự đoán kết quả học tập

Dữ liệu đầu vào cho mạng nơ-ron được đề xuất là gồm 18 thuộc tính với mô tả chi tiết trong Bảng 2.

Bảng 2: Mô tả thuộc tính dữ liệu đầu vào

Stt	Thuộc tính	Mô tả
1	idsv	ID sinh viên
2	f_gioitinh	Giới tính
3	f_matp	Hộ khẩu tỉnh/thành phố
4	f_diemts1	Điểm tuyển sinh môn 1
5	f_diemts2	Điểm tuyển sinh môn 2
6	f_diemts3	Điểm tuyển sinh môn 3
7	idng	Ngành học
8	idkhoa	ID Khoa
9	f_khoahoc	Khóa học
10	f_hocky	Học kỳ
11	f_dtbtl	Điểm trung bình tích lũy đến học kỳ trước
12	f_tdtbtl	Tổng điểm trung bình tích lũy đến học kỳ trước
13	f_tcchiatl	Tổng số tín chỉ tích lũy đến học kỳ trước
14	f_dtbs	Điểm trung bình của học kỳ trước
15	idnv	ID Giảng viên
16	idmh	Môn học
17	f_tinchi	Số tín chỉ môn học
18	f_ngay	Thời gian nhập điểm

Đầu ra của mô hình là điểm môn học của sinh viên theo thang điểm 4, với cách tính cụ thể như Bảng 3.

Bảng 3: Mô tả dữ liệu đầu ra

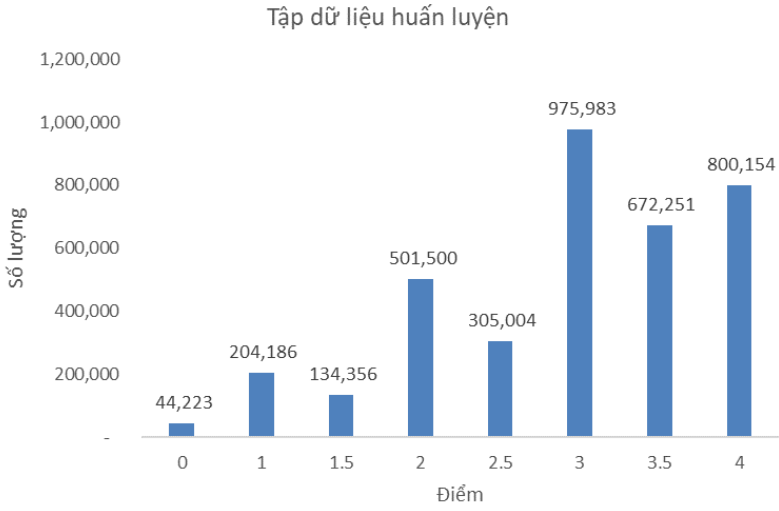
Stt	Điểm chữ	Điểm số (thang điểm 4)
1	A	4,0
2	B+	3,5
3	B	3,0
4	C+	2,5
5	C	2,0
6	D+	1,5
7	D	1,0
8	F	0,0

Kiến trúc chi tiết của mạng MLP gồm 6 tầng. Tầng đầu tiên là tầng input có 18 node (là các thuộc tính trong dữ liệu xem xét), sử dụng hàm kích hoạt là ReLU (rectified linear unit). Tầng ẩn thứ nhất có 256 node, sử dụng hàm kích hoạt là ReLU; tầng ẩn thứ 2 và 3 có 256 node, sử dụng hàm kích hoạt là Sigmoid; tầng ẩn thứ 4 có 256 node, sử dụng hàm kích hoạt là ReLU; tầng ẩn thứ 5 có 8 node, sử dụng hàm kích hoạt là ReLU; tầng thứ 6 là tầng output có 1 node, sử dụng hàm kích hoạt là linear do giá trị đầu ra từ 0 đến 4. Ngoài ra, để tránh vấn đề overfitting khi huấn luyện dữ liệu, chúng tôi sử dụng kỹ thuật dropout với tỷ lệ 0,015 (Srivastava *et al.* 2014). Mạng nơ-ron vốn có số lượng tham số lớn, việc thực hiện dropout sẽ giảm ngẫu nhiên số lượng nơ-ron trong quá trình huấn luyện có thể kéo theo giảm được vấn đề overfitting. Với mạng như hiện nay không quá lớn nên chúng tôi đề xuất dùng tỷ lệ dropout là 0,015 nghĩa là chỉ khoảng 1,15% số lượng nơ-ron sẽ bị bỏ ra ngẫu nhiên trong quá trình huấn luyện để giảm các tham số học qua đó hy vọng giảm được vấn đề overfitting. Bên cạnh kỹ thuật dropout, chúng tôi sử dụng một kỹ thuật khác vừa để hạn chế vấn đề overfitting vừa rút ngắn thời gian huấn luyện mạng là kỹ thuật Early Stopping với giá trị epoch xem xét là 5. Trong trường hợp 5 epoch liên tục độ lỗi không giảm chúng ta sẽ dừng việc huấn luyện. Nếu overfitting không xảy ra, quá trình huấn luyện sẽ chạy tối đa 500 epoch. Batch_size kích thước là 256.

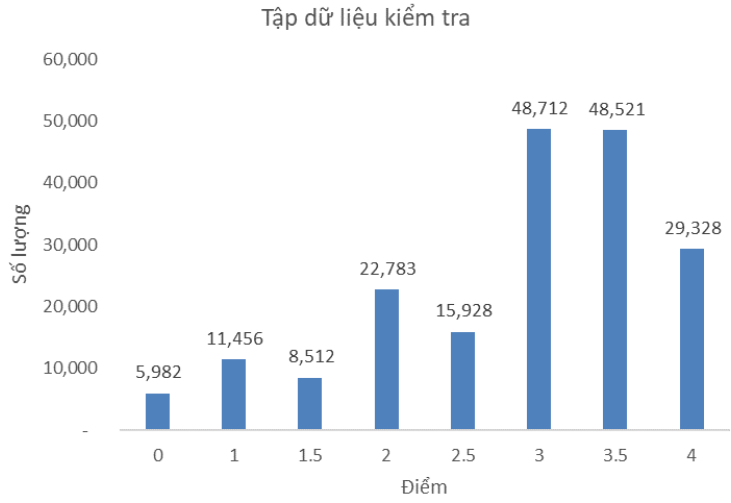
4 ĐÁNH GIÁ KẾT QUẢ

4.1 Dữ liệu huấn luyện và kiểm tra

Dữ liệu thu thập được chia thành hai tập dữ liệu huấn luyện (train) và kiểm tra (test) theo trình tự thời gian. Dữ liệu điểm giai đoạn đầu dùng làm tập train, giai đoạn sau dùng làm tập test. Cụ thể, tập train bao gồm 3.637.657 mẫu tin (tương ứng mỗi mẫu tin là kết quả của một sinh viên cho một môn học nào đó) từ học kỳ 1 năm học 2007-2008 đến học kỳ 1 năm học 2018-2019, chiếm 95% và tập test gồm 191.222 mẫu tin được thu thập ở học kỳ 2 năm học 2018-2019, chiếm 5%. Mục đích việc chia dữ liệu như trên là để bám theo thực tế: dựa trên các môn mà sinh viên đã học để dự đoán cho các môn trong học kỳ tiếp theo. Chi tiết phân phối dữ liệu theo thang điểm từ 0 đến 4 được trình bày trong Hình 5 và Hình 6. Kết quả cho thấy dữ liệu chủ yếu tập trung ở phần điểm trên 3.0, những phần dữ liệu các điểm khác ít hơn. Tỷ lệ thành phần điểm ở 2 phần dữ liệu huấn luyện và dữ liệu để kiểm tra chất lượng mô hình là khá tương đồng.



Hình 5: Phân phối dữ liệu của tập huấn luyện



Hình 6: Phân phối dữ liệu của tập kiểm tra

4.2 Các độ đo dùng để đánh giá

Để đánh giá mô hình, nghiên cứu này sử dụng hai độ đo phổ biến là RMSE (root mean square error) và MAE (mean absolute error).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|Y_j - \hat{Y}_j|$$

Trong đó, Y_i là giá trị dự đoán mẫu thứ i và $\hat{Y}_i$ là giá trị thực tế của mẫu thứ i, n là số mẫu dùng để đánh giá.

4.3 Các Baseline và phương pháp dùng để so sánh

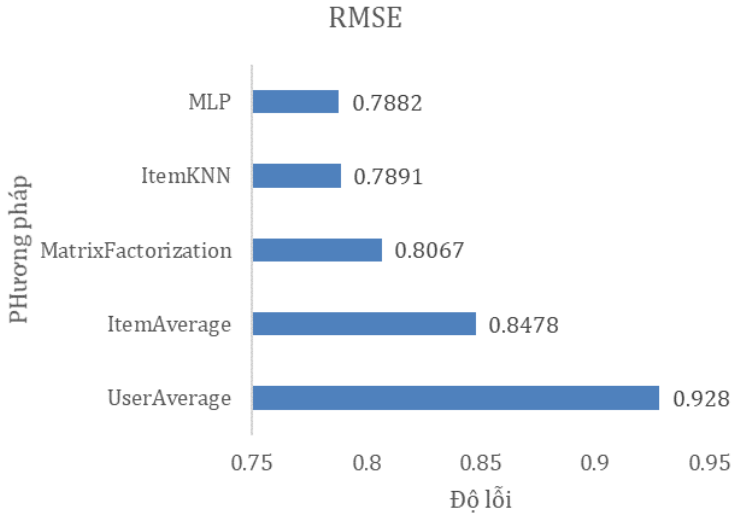
Các baseline đã được sử dụng để so sánh kết quả là User Average (dự đoán dựa trên kết quả trung bình của từng sinh viên), Item Average (dự đoán dựa trên kết quả trung bình của từng môn học). Ngoài ra, chúng tôi cũng so sánh với các phương pháp khác được sử dụng khá thành công trước đây trong dự đoán kết quả học tập là Collaborative Filtering như Item-kNN; chi tiết về các phương pháp này được mô

tả trong bởi (Thai-Nghe and Schmidt-Thieme, 2015; Khanal *et al.*, 2019) và kỹ thuật nổi trội (state-of-the-art) trong lĩnh vực Hệ thống gợi ý là Matrix Factorization (Koren *et al.*, 2009) đã được sử dụng khá thành công trong dự đoán kết quả học tập.

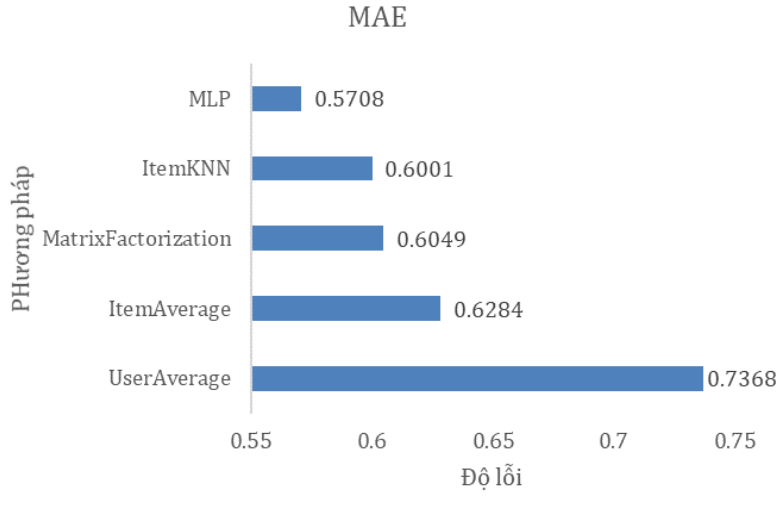
4.4 Kết quả thực nghiệm

Để tránh tính ngẫu nhiên của giải thuật, mỗi phương pháp được thực hiện 10 lần và tính kết quả trung bình độ lỗi trên 10 lần chạy (chi tiết được mô tả trong Hình 7 và Hình 8). Kết quả cho thấy mô hình đề xuất đã cải thiện được độ lỗi so với các phương pháp trước đây sử dụng trong dự đoán kết quả học tập. Kết quả của mô hình MLP cho kết quả tốt nhất ở cả 2 độ đo, trong đó với độ đo MAE, mô hình này cho kết quả vượt trội so với những phương pháp còn lại. Mô hình dựa vào điểm trung bình của sinh viên cho kết quả kém nhất, trong khi mô hình ItemKNN có kết quả gần tương đương với mô hình đề xuất, trong khi kết quả của mô hình Matrix Factorization đứng ở vị trí "trung bình" trong tất cả các phương pháp được khảo sát.

Kết quả cho thấy, việc dự đoán kết quả học tập nhằm xác định và phát hiện sớm các đối tượng sinh viên yếu kém cần được hỗ trợ, tránh việc cảnh báo học vụ và buộc thôi học. Ngoài ra, dự đoán kết quả học tập cũng nhằm xác định được các sinh viên giỏi làm nòng cốt để bồi dưỡng đào tạo, từ đó giúp ích rất nhiều cho bản thân sinh viên, gia đình và xã hội.



Hình 7: Độ lỗi RMSE giữa các phương pháp



Hình 8: Độ lỗi MAE giữa các phương pháp

5 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Bài viết này đề xuất một phương pháp dự báo kết quả học tập của sinh viên bằng kỹ thuật học sâu nhằm khai thác cơ sở dữ liệu trong hệ thống quản lý sinh viên tại các trường đại học. Sau khi thu thập dữ liệu, chúng tôi tiến hành phân tích lựa chọn các thuộc tính phù hợp, tiền xử lý dữ liệu, thiết kế và huấn luyện mạng MLP. Kết quả thực nghiệm cho thấy mô hình đề xuất cho kết quả dự đoán khá chính xác và hoàn toàn khả thi để áp dụng vào thực tế.

Mô hình này có thể tiếp tục cải tiến bằng cách bổ sung các thuộc tính đầu vào như điểm kiểm tra chất lượng Anh văn đầu vào, điểm rèn luyện trong quá trình học tập tại trường,... và hiệu chỉnh mô hình nhằm cải tiến hơn nữa kết quả dự đoán. Một số so sánh thêm với những kỹ thuật tiên tiến khác cũng cần được thực hiện để so sánh với phương pháp đề xuất trong các nghiên cứu tương lai.

LỜI CẢM ƠN

Đề tài này được tài trợ bởi Dự án Nâng cấp Trường Đại học Cần Thơ VN14-P6 bằng nguồn vốn vay ODA từ Chính phủ Nhật Bản.

TÀI LIỆU THAM KHẢO

Altabrawee, H., Ali, O. A. J. and Ajmi, S. Q., 2019. Predicting students' performance using machine learning techniques. Journal of University of Babylon for Pure and Applied Sciences, 27(1): 194-205.

Đại học Cần Thơ, 2020. *Hệ thống thông tin quản lý*, ngày truy cập 12/5/2020. Địa chỉ: https://htql.ctu.edu.vn/.

Fu, M., Qu, H., Yi, Z., Lu, L. and Liu, Y., 2019. A novel deep learning-based collaborative filtering model for recommendation system. IEEE Transactions on Cybernetics. 49(3): 1084-1096.

Guo, B., Zhang, R., Xu, G., Shi, C. and Yang, L., 2015. Predicting Students Performance in Educational Data Mining. 2015 International Symposium on Educational Technology (ISET), pp. 125-128.

Hai, N. P., Sheu, T.-W. and Nagai, M., 2015. Dự báo kết quả học tập của học sinh dựa trên sự kết hợp phương pháp gần đúng Taylor và các mô hình xám. VNU Journal of Science: Education Research. 31 (2): 70-83.

Huynh-Ly, T.-N. and Thai-Nghe, N., 2013. Hệ thống dự đoán kết quả học tập của sinh viên sử dụng thư viện hệ thống gợi ý mã nguồn mở Mymedialite. Hội thảo Quốc gia lần thứ XVI "Một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông", Trường Đại học Cần Thơ.

Iqbal, Z., Qadir, J., Mian, A. and Kamiran, F., 2017. Machine Learning Based Student Grade Prediction: A Case Study. Computers and Society. DOI: arxiv.org/abs/1708.08744.

Khanal, S. S., Prasad, P. W. C., Alsadoon, A. and Maag, A., 2019. A systematic review: machine learning based recommendation systems for e-learning. Education and Information Technologies. DOI: 10.1007/s10639-019-10063-9

Koren, Y., Bell, R. and Volinsky, C., 2009. Matrix factorization techniques for recommender systems. Computer. (8): 30-37.

Nguyen Thai, N., Janecek, P. and Haddawy, P., 2007. A comparative analysis of techniques for predicting academic performance. 2007 37th Annual Frontiers In Education Conference - Global Engineering: Knowledge Without Borders, Opportunities Without Passports, pp. T2G-7-T2G-12.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R., 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of machine learning research. 15: 1929-1958.

Tanuar, E., Heryadi, Y., Lukas, Abbas, B. S. and Gaol, F. L., 2018. Using Machine Learning Techniques to Earlier Predict Student's Performance. 2018 Indonesian Association for Pattern Recognition International Conference (INAPR), pp. 85-89.

Thai-Nghe, N. and Schmidt-Thieme, L., 2015. Factorization forecasting approach for user modeling. Journal of Computer Science and Cybernetics. 31: 133-147.

Thai-Nghe, N., Horvath, T. and Schmidt-Thieme, L., 2011. Factorization Models for Forecasting Student Performance. Proceedings of the 4th International Conference on Educational Data Mining. Eindhoven, The Netherlands, pp. 11-20.

Zhang, L., Luo, T., Zhang, F. and Wu, Y., 2018. A Recommendation Model Based on Deep Neural Network. IEEE Access. 6: 9454-9463.