

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC LẠC HỒNG



HUỲNH LÝ THANH NHÀN

GIẢI PHÁP GỢI Ý TRONG VIỆC CỐ VẤN HỌC TẬP

TÓM TẮT LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

Ngành: Khoa học máy tính

Mã số ngành: **9480101**

Đồng Nai – năm 2024

Công trình được hoàn thành tại: Trường Đại học Lạc Hồng

Người hướng dẫn khoa học:

PGS.TS. Nguyễn Thái Nghe

PGS.TS. Lê Huy Thập

Phản biện 1:

Phản biện 2:

Phản biện 3:

Luận án sẽ được bảo vệ trước Hội đồng chấm luận án cấp Trường họp tại

.....

.....

Vào hồi giờ....., ngày.....tháng.....năm....

Có thể tìm hiểu luận án tại thư viện:

- Thư viện trường Đại học Lạc Hồng
- Thư viện Quốc Gia

MỤC LỤC

CHƯƠNG 1: TỔNG QUAN	1
1.1 Tổng quan về đề tài luận án.....	1
1.2 Bài toán nghiên cứu và ý nghĩa	1
1.2.1 Sự tương quan giữa bài toán xếp hạng và dự đoán kết quả học tập.....	1
1.2.2 Bài toán dự đoán kết quả học tập của sinh viên.....	2
1.2.3 Gợi ý lựa chọn môn học tự chọn	3
1.3 Thách thức của bài toán nghiên cứu	4
1.4 Các vấn đề nghiên cứu.....	5
1.5 Mục tiêu nghiên cứu và hướng tiếp cận của luận án	5
1.6 Đối tượng và phạm vi nghiên cứu	6
1.7 Các đóng góp của luận án.....	6
1.8 Cấu trúc của luận án	6
CHƯƠNG 2: CƠ SỞ LÝ THUYẾT VÀ NGHIÊN CỨU LIÊN QUAN.....	7
2.1 Giới thiệu về hệ trợ giảng thông minh.....	7
2.2 Giới thiệu về hệ thống gợi ý	7
2.3 Khai thác dữ liệu giáo dục	7
2.4 Các nghiên cứu liên quan	7
2.5 Tổng kết chương và thảo luận	8
CHƯƠNG 3: DỰ ĐOÁN KẾT QUẢ HỌC TẬP CỦA SINH VIÊN THEO HƯỚNG TIẾP CẬN HỆ THỐNG GỢI Ý	8
3.1 Giải bài toán dự đoán kết quả học tập sinh viên theo hướng tiếp cận hệ thống gợi ý.....	8
3.2 Phương pháp lọc cộng tác theo sinh viên tương tự (Student-kNNs)	8

3.3 Phương pháp lọc cộng tác theo môn học tương tự (Course-kNNs).....	9
3.4 Phương pháp phân rã ma trận - Matrix Factorization.....	9
3.5 Phương pháp phân rã ma trận thiên vị - Biased Matrix Factorization	10
3.6 Đánh giá kết quả.....	11
3.7 Tổng kết chương.....	12
CHƯƠNG 4: CÁC MÔ HÌNH PHÂN RÃ SÂU MA TRẬN ĐỂ DỰ ĐOÁN KẾT QUẢ HỌC TẬP CỦA SINH VIÊN.....	12
4.1 Đặt vấn đề và phương pháp giải quyết	12
4.2 Phương pháp tích hợp kiến trúc học sâu vào phân rã ma trận	13
4.3 Phương pháp phân rã sâu ma trận thiên vị.....	16
4.4 Đánh giá kết quả	16
4.5 Tổng kết chương.....	17
CHƯƠNG 5: TÍCH HỢP CÁC MỐI QUAN HỆ DỮ LIỆU TRONG DỰ ĐOÁN KẾT QUẢ HỌC TẬP	18
5.1 Đặt vấn đề và các phương pháp giải quyết.....	18
5.2 Phương pháp tích hợp mối quan hệ của sinh viên	19
5.3 Phương pháp tích hợp mối liên quan giữa các môn học.....	20
5.4 Đánh giá kết quả	21
5.5 Tổng kết chương.....	22
CHƯƠNG 6: KẾT LUẬN VÀ KIẾN NGHỊ.....	23
6.1 Kết luận.....	23
6.2 Hạn chế và hướng phát triển.....	24

CHƯƠNG 1: TỔNG QUAN

1.1 Tổng quan về đề tài luận án

Thời gian gần đây, số lượng sinh viên bị buộc thôi học có chiều hướng tăng ở nhiều trường đại học và thường tập trung vào những sinh viên học năm thứ ba và năm thứ tư. Một phần nguyên nhân là do sinh viên không có kế hoạch học tập phù hợp. Chính vì thế việc phát hiện sớm năng lực của từng sinh viên để giúp họ lập kế hoạch học tập sao cho phù hợp là một trong những nghiên cứu hỗ trợ cho cố vấn học tập và có nhu cầu rất cần thiết [1].

Để đánh giá năng lực học tập của sinh viên nhằm đưa ra những gợi ý lựa chọn môn học hợp lý. Nhiều nghiên cứu đã xuất bản về dự đoán năng lực học tập sinh viên từ các phương pháp kỹ thuật phổ biến như mô hình mạng Bayes và cây quyết định [2][3], luật kết hợp và luật kết hợp dựa trên chuỗi [4][5], giải thuật di truyền, lý thuyết trò chơi [6], hoặc sử dụng lý thuyết tập thô [7], và kể cả theo xu thế hiện nay như mô hình học sâu [8][9], hoặc dùng phương pháp lập luận theo tình huống (Case-based reasoning - CBR) [10].

1.2 Bài toán nghiên cứu và ý nghĩa

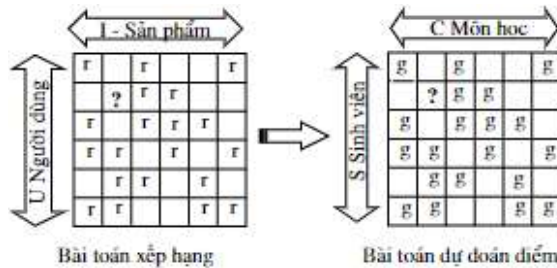
1.2.1 Sự tương quan giữa bài toán xếp hạng và dự đoán kết quả học tập

Gần đây việc áp dụng (Recommender System –RS) vào dự đoán kết quả học tập của sinh viên cũng được đầu tư nghiên cứu và phát triển bởi sự tương đồng giữa bài toán dự đoán kết quả học tập của sinh viên và bài toán xếp hạng trong hệ thống gợi ý [11]. Ví dụ, sinh viên học tập các môn học sẽ có điểm số, người dùng mua sản phẩm sẽ có đánh giá sản phẩm (cho từ 1 đến 5 sao). Hình 1.1 thể hiện việc tương đồng giữa bài toán dự đoán kết quả học tập và bài toán xếp hạng sản phẩm.



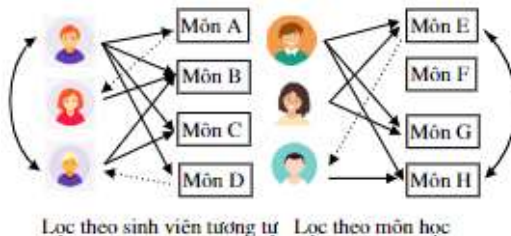
Hình 1.1: Sự tương đồng giữa hệ thống RS và hệ thống dự đoán kết quả học tập

Trong hệ thống gợi ý được cấu thành từ danh sách người dùng (user- u), danh sách các đối tượng như bài hát, bộ phim, sản phẩm (item – i) và các đánh giá (ratings - r) là chỉ số đánh giá của người dùng u trên đối tượng i . Tương tự, trong bài toán dự đoán điểm học tập của sinh viên thì có danh sách sinh viên s , danh sách môn học c và điểm g . Như vậy, việc dự đoán đánh giá của người dùng trong bài toán xếp hạng (rating prediction) của RS sẽ tương đương với bài toán dự đoán điểm sinh viên (Hình 1.2). Sự ánh xạ này được biểu diễn: {Người dùng $\rightarrow$ Sinh viên}, { Sản phẩm $\rightarrow$ Môn học }, {Điểm $\rightarrow$ Đánh giá}.



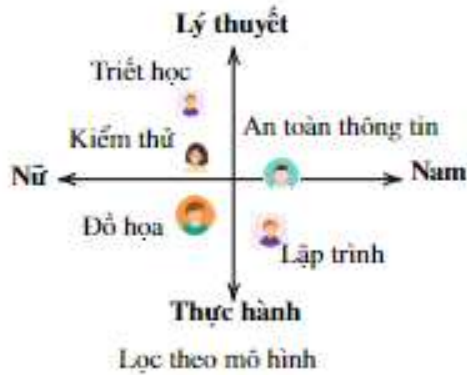
Hình 1.2: Sự tương đồng giữa bài toán dự đoán điểm với bài toán xếp hạng
1.2.2 Bài toán dự đoán kết quả học tập của sinh viên

Mô hình dự đoán năng lực học tập sinh viên theo hướng tiếp cận cá nhân hóa dựa trên ý tưởng của các phương pháp lọc cộng tác (Colaborative Filtering - CF) trong hệ thống gợi ý. Bên dưới là minh họa bài toán dự đoán năng lực sinh viên theo ba nhóm: Student-based filtering và Course-based filtering (Hình 1.3), và Model-based filtering (Hình 1.4). Student-based filtering sẽ dựa vào năng lực tương đồng của 2 sinh viên và kết quả của một sinh viên đã học mà dự đoán kết quả cho sinh viên chưa học. Tương tự, Course-based filtering sẽ dựa vào môn học giống nhau mà dự đoán cho sinh viên đã học môn tương tự.



Hình 1.3: Các phương pháp lọc năng lực tương đồng

Model-based filtering dựa mô hình nhân tố tiềm ẩn mà dự đoán kết quả các môn học cho từng sinh viên cụ thể.



Hình 1.4: Mô hình nhân tố tiềm ẩn dự đoán kết quả học tập sinh viên

1.2.3 Gợi ý lựa chọn môn học tự chọn

Trước khi sinh viên lập kế hoạch học tập theo từng học kỳ, ta khai phá dữ liệu điểm số của sinh viên để đưa ra kết quả dự đoán. Kết quả dự đoán này làm cơ sở cho hệ thống đưa ra gợi ý nên chọn môn học nào trong những môn tự chọn là phù hợp mà vẫn đảm bảo các ràng buộc của chương trình đào tạo.

Ví dụ: Có 05 sinh viên: sv1, sv2, sv3, sv4 và sv5 học các môn Môn1, Môn2, ..., Môn n, Môn n1, Môn n2, Môn n3 được trình bày trong một ma trận như Hình 1.5, mỗi ô trong ma trận chứa số điểm của sinh viên học môn học tương ứng, những sinh viên chưa học môn nào thì sẽ điền giá trị ô đó bởi dấu chấm hỏi "?". Trong những môn học đó có 3 môn học tự chọn là Môn n1, Môn n2, Môn n3. Sinh viên cần chọn 2 môn trong 3 môn học tự chọn sao cho có kết quả phù hợp với mình nhất. Như vậy, hệ thống cần gợi ý cho sinh viên sv5 là nên học 2 môn nào trong 3 môn: Môn n1, Môn n2, Môn n3.

Sau khi chạy giải thuật dự đoán cho tất cả sinh viên học tất cả các môn học mà sinh viên đó chưa học và điền kết quả vào ma trận. Từ những ràng buộc về số tín chỉ hay số môn học tự chọn mà sinh viên cần học trong một học kỳ để đưa ra gợi ý phù hợp.

		C - Course (Môn học)						
		Môn 1	Môn 2		Môn n	Môn n1	Môn n2	Môn n3
S - Student (Sinh viên)	SV 1	2	2		2	1	4	?
	SV 2	3	2		4	?	2	2
	SV 3	1	2		?	3	?	1
	SV 4	2	3		2	3	?	2
	SV 5	1	4		4	?	?	?

Hình 1.5: Dữ liệu điểm với ba môn cần dự đoán của sinh viên sv5

Trở lại ví dụ trên, hệ thống cần gợi ý 2 môn học tự chọn cho sinh viên sv5 là 2 môn: Môn n1 và Môn n2. Vì 2 môn học này có số điểm dự đoán cao hơn môn học Môn n3 (3 và 4 > 2) như Hình 1.6.

		C - Course (Môn học)						
		Môn 1	Môn 2		Môn n	Môn n1	Môn n2	Môn n3
S - Student (Sinh viên)	SV 1	2	2		2	1	4	2
	SV 2	3	2		4	3	2	2
	SV 3	1	2		2	3	3	1
	SV 4	2	3		2	3	3	2
	SV 5	1	4		4	3	4	2

Hình 1.6: Bảng điểm sau khi dự đoán và hướng gợi ý

1.3 Thách thức của bài toán nghiên cứu

Dự đoán theo hướng cá nhân hóa: Mỗi người học có năng lực khác nhau, hoàn cảnh khác nhau, nên ta cần đưa ra dự đoán riêng cho từng sinh viên, trong khi phần lớn các nghiên cứu hiện nay chỉ tập trung vào các quy luật chung.

Đánh giá không đồng nhất (Biased Ratings): Trong ngữ cảnh hệ thống gợi ý, nhiều đánh giá chịu ảnh hưởng từ phía người dùng (user) hoặc đối tượng (item), ảnh hưởng này được gọi là thiên vị / thiên hướng (bias) và tồn tại một cách độc lập với quan hệ chủ quan giữa người dùng và đối tượng. Thiên vị này được thể hiện (trong dữ liệu) xu hướng có tính hệ thống, trong đó một số người dùng sẽ cho điểm đánh giá cao (hoặc thấp) hơn số khác hoặc một số đối tượng

được chấp nhận tốt hơn (hoặc kém hơn) so với các đối tượng khác. Tương tự như vậy, trong ngữ cảnh giáo dục, có những đánh giá (chấm điểm) không đồng nhất giữa các giảng viên khác nhau (có những giảng viên chấm điểm thường cao hoặc thường thấp hơn), cũng như những yêu cầu khó hoặc dễ khác nhau của các môn học, vấn đề này sẽ ảnh hưởng đến kết quả dự đoán và hiệu quả của mô hình dự đoán, cho nên cần được khắc phục.

Thiếu xem xét các mối quan hệ người dùng và mối quan hệ giữa các môn học. Các sinh viên có cùng môi trường sẽ ảnh hưởng lên nhau, cũng như những môn học có nội dung liên quan nhau, đây là điểm thuận lợi cho ta khai thác.

Độ chính xác của dự đoán còn có thể nâng cao bằng các mô hình tích hợp. Hiện nay, khoa học càng phát triển, con người luôn tìm được những điểm mới đã thành công ở lĩnh vực này, thì có thể áp dụng thành công ở lĩnh vực khác.

1.4 Các vấn đề nghiên cứu

Từ việc phân tích các thách thức, vấn đề còn bỏ ngõ của các nghiên cứu liên quan, luận án xác định một số vấn đề nghiên cứu bao gồm: vấn đề dự đoán năng lực học tập của sinh viên theo hướng cá nhân hóa, vấn đề không đồng nhất trong đánh giá năng lực học tập của sinh viên, vấn đề xét mối quan hệ bạn bè của sinh viên có thể nâng cao được hiệu quả dự đoán, vấn đề xét mối liên quan giữa các môn học, và vấn đề cải thiện độ chính xác của dự đoán bằng cách tích hợp các mô hình tiên tiến.

1.5 Mục tiêu nghiên cứu và hướng tiếp cận của luận án

Mục tiêu chính của luận án là đề xuất các phương pháp tiếp cận mới cho “bài toán dự đoán kết quả học tập của sinh viên” từ đó xây dựng hệ thống gợi ý lựa chọn môn học phù hợp cho sinh viên.

Mục tiêu 1: Đề xuất các kỹ thuật khai phá dữ liệu dự đoán năng lực học tập của sinh viên để đưa gợi ý lựa chọn môn học theo hướng cá nhân hóa.

Mục tiêu 2: Đề xuất các phương pháp giải quyết vấn đề đánh giá không đồng đều (đã nêu trong thách thức nghiên cứu thứ hai).

Mục tiêu 3: Đề xuất các phương pháp tích hợp mối quan hệ giữa các sinh viên và mối liên quan giữa các môn học vào mô hình dự đoán nhằm nâng cao độ chính xác trong dự đoán kết quả học tập của sinh viên.

Mục tiêu 4: Đề xuất các phương pháp tích hợp mối liên quan giữa các môn học vào mô hình dự đoán nhằm nâng cao độ chính xác trong dự đoán kết quả học tập của sinh viên.

Mục tiêu 5: Đề xuất ứng dụng ý tưởng của kỹ thuật học sâu vào mô hình phân rã ma trận và phân rã ma trận thiên vị nhằm nâng cao hiệu quả dự đoán.

1.6 Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu: các mô hình dự đoán và các dữ liệu khai phá.

Phạm vi nghiên cứu của luận án: là tập trung vào các phương pháp xây dựng mô hình dự đoán, thực nghiệm trên 2 tập dữ liệu ở Việt Nam và Quốc tế.

1.7 Các đóng góp của luận án

Đóng góp thứ nhất của luận án là đề xuất mô hình gợi ý lựa chọn môn học cho sinh viên theo hướng tiếp cận hệ thống gợi ý, đồng thời vấn đề dữ liệu đánh giá không đồng nhất cũng được giải quyết (công bố [CT1]).

Đóng góp thứ hai của luận án là đề xuất mô hình tích hợp tăng cường kiến trúc học sâu vào giải thuật phân rã ma trận cũng như giải thuật phân rã ma trận thiên vị nhằm nâng cao hiệu quả dự đoán (công bố [CT4], [CT5]).

Đóng góp thứ ba của luận án là đề xuất phương pháp tích hợp mối quan hệ người dùng vào mô hình dự đoán điểm sinh viên nhằm tận dụng được sự ảnh hưởng của mối quan hệ bạn bè cùng lớp lên kết quả học tập để từ đó có thể nâng cao được hiệu quả dự đoán (công bố [CT2]).

Đóng góp thứ tư là luận án là đề xuất các phương pháp tích hợp sử dụng dữ liệu bổ sung thông tin như mối liên quan giữa các môn học (công bố [CT3]).

Sau quá trình nghiên cứu, thực hiện luận án, tác giả đã công bố được 02 bài báo đăng trong kỷ yếu hội thảo quốc tế (IEEE/ Scopus indexed), 01 bài báo đăng trong kỷ yếu hội thảo quốc tế (Springer /Scopus indexed) và 02 bài báo đăng trên tạp chí quốc tế (Scopus indexed 2020 và DBLP indexed).

1.8 Cấu trúc của luận án

Luận án được cấu trúc như sau: tóm tắt, chương 1 – giới thiệu, chương 2 – tổng quan tình hình nghiên cứu, chương 3,4&5 – các bài toán nghiên cứu chính, chương 6 – kết luận, danh mục các bài báo đã công bố.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT VÀ NGHIÊN CỨU LIÊN QUAN

2.1 Giới thiệu về hệ trợ giảng thông minh

Hệ trợ giảng thông minh (Intelligent Tutoring System - ITS) là một hệ thống có khả năng tùy chỉnh và phản hồi kết quả tức thì cho người học mà không cần sự can thiệp hay trợ giúp của giáo viên [12]. ITS được cấu thành từ ba lĩnh vực khác nhau: Khoa học máy tính, giáo dục học, và tâm lý học.

2.2 Giới thiệu về hệ thống gợi ý

Hệ thống gợi ý (Recommender Systems - RS) là một dạng của hệ thống lọc thông tin, nó được sử dụng để dự đoán sở thích hay xếp hạng mà người dùng có thể dành cho một mục thông tin nào đó mà họ chưa xem xét tới trong quá khứ (mục thông tin có thể là bài hát, bộ phim, đoạn video clip, sách, sản phẩm, bài báo, tin tức..).

2.3 Khai thác dữ liệu giáo dục

Dự đoán kết quả học tập của sinh viên một cách chính xác là rất hữu ích trong nhiều ngữ cảnh khác nhau ở các trường đại học. Chẳng hạn, Cố vấn học tập có thể xác định các ứng viên xuất sắc để đề cử tham gia các cuộc thi tài năng, hoặc cấp học bổng nhằm khuyến khích họ nỗ lực hơn nữa trong học tập, hay việc xác định các sinh viên có năng lực yếu kém (có nguy cơ bị buộc thôi học) để có những biện pháp thích hợp nhằm hỗ trợ họ học tập tốt hơn, hoặc đơn giản là gợi ý lựa chọn các môn học tự chọn trong học chế tín chỉ.

2.4 Các nghiên cứu liên quan

Do tính chất cấp thiết của việc đánh giá năng lực học tập của sinh viên nhằm đưa ra những gợi ý tư vấn hợp lý. Nhiều nghiên cứu đã xuất bản về dự đoán năng lực học tập sinh viên từ các phương pháp kỹ thuật truyền thống như cây quyết định, hồi qui tuyến tính, máy học véc-tơ hỗ trợ SVM, mô hình mạng Bayes [13], luật kết hợp tuần tự [4], giải thuật di truyền [6] hoặc sử dụng lý thuyết tập thô [7], ứng dụng RS vào khai phá dữ liệu giáo dục và nhiều nghiên cứu khác về khai phá dữ liệu giáo dục [14][15] [16] [17]. Tuy nhiên, kết quả đo độ lỗi RMSE vẫn còn cao và cần nghiên cứu nâng cao độ chính xác cho các giải thuật dự đoán.

2.5 Tổng kết chương và thảo luận

Từ những cơ sở lý thuyết và kết quả của các nghiên cứu liên quan, luận án đã đưa ra cái nhìn tổng quan về bài toán giải pháp gợi ý trong việc cố vấn học tập của sinh viên.

CHƯƠNG 3: DỰ ĐOÁN KẾT QUẢ HỌC TẬP CỦA SINH VIÊN THEO HƯỚNG TIẾP CẬN HỆ THỐNG GỢI Ý

Nội dung chương là đề xuất các mô hình dự đoán kết quả học tập sinh viên theo hướng tiếp cận hệ thống gợi ý như: lọc công tác theo sinh viên tương tự, môn học tương tự, đặc biệt là hai kỹ thuật phân rã ma trận cũng như phân rã ma trận thiên vị. Qua kết quả thực nghiệm, cho thấy các phương pháp đề xuất có độ chính xác tốt hơn các phương pháp cơ sở, nội dung của chương này được tổng hợp từ công bố [CT1].

3.1 Giải bài toán dự đoán kết quả học tập sinh viên theo hướng tiếp cận hệ thống gợi ý

Các bài toán khác trong lĩnh vực khai thác dữ liệu, xây dựng hệ thống dự đoán kết quả học tập của sinh viên cũng tuân theo quy trình chuẩn CRISP-DM (CRoss Industry Standard Process for Data Mining). Quy trình này được thiết kế để giải quyết bài toán khai thác dữ liệu và gồm sáu giai đoạn, tương tự như mô hình thác đổ trong phân tích và thiết kế hệ thống thông tin, bao gồm: Tìm hiểu vấn đề, tìm hiểu dữ liệu, chuẩn bị dữ liệu, mô hình hóa, đánh giá mô hình và triển khai ứng dụng.

3.2 Phương pháp lọc cộng tác theo sinh viên tương tự (Student-kNNs)

Phương pháp lọc cộng tác dựa trên người dùng [18] được xây dựng trên ý tưởng sử dụng toàn bộ ma trận đánh giá để lựa chọn một tập hợp người dùng có sở thích tương tự nhất với người dùng hiện tại. Bằng cách kết hợp các đánh giá từ tập người dùng tương tự để dự đoán đánh giá cho người dùng hiện tại trên các mục chưa được đánh giá. Tương tự, trong lĩnh vực giáo dục, chúng ta cũng giả định rằng những sinh viên tương tự nhau sẽ có năng lực tương tự trên những môn học tương tự. Thật vậy, phương pháp lọc cộng tác dựa trên người

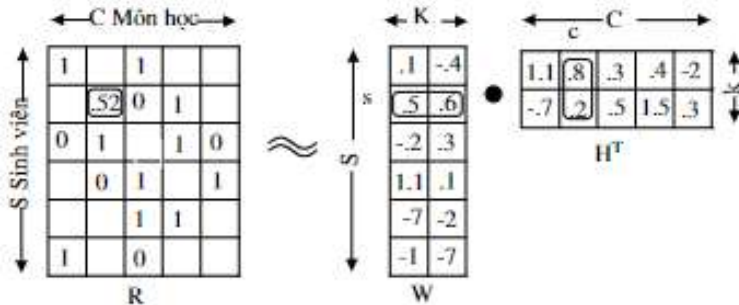
dùng và món hàng sẽ là lựa chọn để xem xét cho bài toán dự đoán năng lực học tập của sinh viên. Phương pháp lọc cộng tác sử dụng láng giềng gần nhất (k-nearest neighbors) trong ngữ cảnh giáo dục được gọi là Student k-NNs

3.3 Phương pháp lọc cộng tác theo môn học tương tự (Course-kNNs)

Phương pháp lọc dựa vào môn học tương tự này khác với phương pháp lọc cộng tác dựa trên sinh viên tương tự, bởi thay vì tính toán mức độ tương tự giữa các sinh viên trong hệ thống với sinh viên hiện tại s , phương pháp này tính toán mức độ tương tự giữa môn học cần dự đoán và các môn học được sinh viên s đã học. Để tính toán mức độ tương tự giữa hai môn học, phương pháp xem xét tập sinh viên đã học cả hai môn học đó. Sau đó, phương pháp chọn ra một tập các môn học láng giềng với môn học cần dự đoán điểm số bởi s . Cuối cùng, phương pháp kết hợp các điểm số của s với tập môn học láng giềng này để đưa ra dự đoán điểm của s với môn học cần dự đoán.

3.4 Phương pháp phân rã ma trận - Matrix Factorization

Kỹ thuật phân rã ma trận là việc chia một ma trận lớn R thành hai ma trận W và H , hai ma trận này có kích thước nhỏ hơn rất nhiều so với ma trận R , sao cho R có thể được xây dựng lại từ hai ma trận nhỏ hơn này càng chính xác càng tốt [19] được minh họa trong (Hình 3.1), nghĩa là $R \approx WH^T$



Hình 3.1: Minh họa kỹ thuật phân rã ma trận

$W \in \mathbb{R}^{|S| \times |K|}$ là một ma trận mà ở đó mỗi dòng s là một véc-tơ bao gồm k nhân tố tiềm ẩn (latent factors) mô tả cho sinh viên s , và $H \in \mathbb{R}^{|C| \times |K|}$ là một ma trận mà ở đó mỗi dòng c là một véc-tơ bao gồm k nhân tố tiềm ẩn mô tả cho môn học c .

Gọi w và h là các phần tử tương ứng của hai ma trận W và H , hay w và h là các véc-tơ bao gồm k nhân tố tiềm ẩn mô tả cho sinh viên s và môn học c , khi đó điểm số g của sinh viên s trên môn học c được dự đoán bởi công thức:

$$\widehat{g}_{sc} = \sum_{k=1}^K w_{sk} h_{ck} = w_s h_c^T \quad (3.1)$$

W và H là các tham số mô hình (còn gọi là các ma trận nhân tố tiềm ẩn) mà chúng ta cần phải xác định bằng cách tối ưu hóa tối thiểu (min) hàm mục tiêu (3.2) theo điều kiện nào đó, chẳng hạn RMSE (Root Mean Squared Error).

$$RMSE = \sqrt{\frac{1}{|D^{test}|_{s,c,g \in D^{test}}} \sum (g_{sc} - \widehat{g}_{sc})^2} \quad (3.2)$$

Tối ưu hóa hàm mục tiêu theo phương pháp giảm dốc ngẫu nhiên (Stochastic Gradient Descent - SGD) [3]

$$o^{MF} = \sum_{(s,c,g) \in D^{train}} (g_{sc} - \sum_{k=1}^K \widehat{w}_{sk} \widehat{h}_{ck})^2 + \lambda (\|W\|_F^2 + \|H\|_F^2) \quad (3.3)$$

Với λ là hệ số chính tắc hóa ($0 \leq \lambda < 1$) và $\|\cdot\|_F^2$ là chuẩn Frobenius. Đại lượng $\lambda \cdot (\|W\|_F^2 + \|H\|_F^2)$ dùng để ngăn ngừa sự quá khớp (over-fitting).

Với hàm mục tiêu mới thì các giá trị w và h sẽ được cập nhật lại theo hai công thức sau:

$$w' = w_{sk} + \beta(2e_{sc}h_{ck} - \lambda w_{sk}) \quad (3.4)$$

$$h' = h_{ck} + \beta(2e_{sc}w_{sk} - \lambda h_{ck}) \quad (3.5)$$

Trong đó: $e_{sc} = g_{sc} - \widehat{g}_{sc}$ là độ lỗi của dự đoán, β là tốc độ học.

Sau quá trình huấn luyện ta được hai ma trận W và H đã tối ưu, thì quá trình dự đoán được thực hiện. Quá trình dự đoán được tính như sau:

$$\widehat{g}_{sc} = \sum_{k=1}^K w_{sk} h_{ck} = w_s h_c^T \quad (3.6)$$

3.5 Phương pháp phân rã ma trận thiên vị - Biased Matrix Factorization

Để giải quyết vấn đề đánh giá khách quan, giảm bớt sự chênh lệch giữa những yêu cầu cao thấp khác nhau của các môn học. Cũng như giảm thiểu sự gợi ý sai lệch do nhìn nhận từ những sinh viên có sở trường hay sở đoản đối với môn học nào đó. Nghiên cứu sử dụng giải thuật Matrix Factorization kết hợp một lượng giá trị đo độ lệch bias để được giải thuật Biased Matrix Factorization (BMF), phương pháp được công bố trong công trình [CT4]. Để dự đoán điểm của sinh viên s cho môn học i được biểu diễn với công thức sau:

$$\hat{g}_{sc} = \mu + b_s + b_c + \sum_{k=1}^K w_{sk} h_{ck} \quad (3.7)$$

Với giá trị μ là giá trị trung bình toàn cục, là năng lực trung bình của tất cả các sinh viên trên tất cả các môn học với tập dữ liệu huấn luyện.

$$\mu = \frac{\sum_{(s,c,g) \in D^{train}} (g - \mu)}{|D^{train}|} \quad (3.8)$$

Giá trị b_s là độ lệch của sinh viên (là giá trị lệch trung bình của năng lực một sinh viên so với giá trị trung bình toàn cục).

$$b_s = \frac{\sum_{(s',c,g) \in D^{train} | s'=s} (g - \mu)}{|\{(s',c,g) \in D^{train} | s' = s\}|} \quad (3.9)$$

Giá trị b_i là độ lệch của môn học (là giá trị lệch trung bình của yêu cầu môn học so với giá trị trung bình toàn cục).

$$b_c = \frac{\sum_{(s,c',g) \in D^{train} | c'=c} (g - \mu)}{|\{(s,c',g) \in D^{train} | c' = c\}|} \quad (3.10)$$

Do có thay đổi giá trị lệch của sinh viên và môn học, nên hàm mục tiêu cũng thay đổi theo công thức sau:

$$o^{BMF} = \sum_{(s,c,g) \in D^{train}} (g_{sc} - \mu - b_s - b_c - \sum_{k=1}^K w_{sk} h_{ck})^2 + \lambda (\|W\|_F^2 + \|H\|_F^2 + b_s^2 + b_c^2) \quad (3.11)$$

Sau quá trình huấn luyện ta được hai ma trận W và H đã tối ưu, thì quá trình dự đoán được thực hiện.

3.6 Đánh giá kết quả

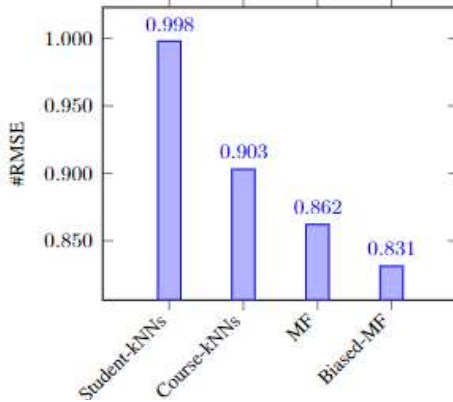
Tập dữ liệu - Tập dữ liệu ASSISTments được xuất bản bởi Nền tảng ASSISTments. Sau khi tiền xử lý, tập dữ liệu này chứa 8519 sinh viên (người dùng), 35978 nhiệm vụ (mục) và 1011079 điểm (xếp hạng).

Đánh giá kết quả - Để đánh giá giải thuật, ta sử dụng nghi thức kiểm tra hold-out: lấy ngẫu nhiên 2/3 tập dữ liệu để làm tập học và 1/3 còn lại để làm tập kiểm tra. Do bài toán dự đoán kết quả học tập của sinh viên thuộc dạng rating prediction (dự đoán từ đánh giá tường minh) thì cách đánh giá phù hợp nhất là Root Mean Squared Error (RMSE). Hơn nữa, các giải thưởng lớn trong lĩnh vực RS đều dùng RMSE để đánh giá, như Netflix Prize, KDD Cup 2010,...

Tìm kiếm siêu tham số (hyper-parameters) - Độ chính xác dự đoán phụ thuộc vào các tham số cung cấp cho thuật toán. Nếu các tham số không phù hợp, độ chính xác dự đoán sẽ không tốt mặc dù thuật toán là đúng. Vì vậy,

việc tìm kiếm tham số tốt nhất là rất quan trọng. Tìm kiếm siêu tham số gồm giai đoạn: tìm thô (Coarse Search) và tìm mịn (Granularity Search).

Bốn thí nghiệm được tiến hành và kết quả thử nghiệm được hiển thị trong (Hình 3.2). So sánh độ lỗi RMSE của các phương pháp khác nhau, biểu đồ cho thấy độ lỗi của tiếp cận được đề xuất là nhỏ nhất trên tập dữ liệu CTU.



Hình 3.2: So sánh độ lỗi RMSE của các phương pháp

3.7 Tổng kết chương

Chương này đã giới thiệu một cách tiếp cận sử dụng phương pháp MF nhằm dự đoán kết quả học tập của sinh viên. Tiến hành thử nghiệm trên tập dữ liệu chuẩn cho thấy rằng quy trình đề xuất hoạt động tốt. Việc áp dụng phương pháp đã cho kết quả khả quan. Tuy nhiên, độ chính xác của dự đoán còn có thể nâng cao bởi các phương pháp cải tiến sẽ được trình bày ở Chương 4 tiếp theo.

CHƯƠNG 4: CÁC MÔ HÌNH PHÂN RÃ SÂU MA TRẬN ĐỂ DỰ ĐOÁN KẾT QUẢ HỌC TẬP CỦA SINH VIÊN

Nội dung chương là đề xuất tích hợp mô hình học sâu vào kỹ thuật phân rã ma trận đã nâng cao độ chính xác cho dự đoán và hoàn toàn thích hợp để xây dựng ứng dụng vào hệ thống gợi ý môn học trong cố vấn học tập [CT4].

4.1 Đặt vấn đề và phương pháp giải quyết

Gần đây, việc ứng dụng học sâu (Deep Learning – DL) vào khai phá dữ liệu giáo dục cũng ngày được quan tâm. Việc kết hợp DL vào kỹ thuật phân rã

ma trận sẽ hứa hẹn nhiều kết quả khả quan. Ngày nay dữ liệu giáo dục càng được thu thập nhiều hơn và có khả năng ứng dụng những kỹ thuật dữ liệu lớn.

4.2 Phương pháp tích hợp kiến trúc học sâu vào phân rã ma trận

Ý tưởng chính của việc tích hợp kiến trúc học sâu vào kỹ thuật MF là tái lập việc phân rã ma trận, lặp lại gần đúng ma trận đầu ra cho đến khi đạt kết quả tốt nhất. Mỗi bước của quy trình là phương pháp cơ sở (MF); mô hình hoàn chỉnh là tổng các ma trận được lưu trên ngăn xếp của các bước đã thực hiện.

Hình 4.1 minh họa hoạt động của mô hình (Deep Matrix Factorization - DMF). Mô hình được khởi tạo với đầu vào là các tham số cơ bản của MF: một ma trận điểm số X của sinh viên đối với môn học nào đó. Ma trận $X = G^0$, là điểm bắt đầu của quá trình. Xấp xỉ ma trận $X \in G^{|S| \times |C|}$ theo tích của hai ma trận nhỏ hơn, W^0 và H^0 , có dạng $\hat{G}^0 = W^0 \cdot H^0$. Ma trận $\hat{G}^0$ cung cấp tất cả các điểm dự đoán và được lưu trữ trong ngăn xếp ở bước đầu tiên. Tại bước này, đệ quy bắt đầu. Ma trận mới $G^1 = X - \hat{G}^0$ được xây dựng bằng cách tính toán các sai số mắc phải giữa ma trận phân loại ban đầu X và các phân loại dự đoán được lưu trữ trong $\hat{G}^0$. Phép phân rã lại xấp xỉ ma trận mới G^1 này thành hai ma trận hạng nhỏ mới $\hat{G}^1 = W^1 \cdot H^1$, tạo ra sai số ở bước thứ hai $G^2 = G^1 - \hat{G}^1$. Quá trình này được lặp lại nhiều lần bằng cách tạo và phân rã các ma trận lỗi liên tiếp $G^1, \dots, G^T$. Ma trận này sẽ hội tụ về 0. Tương tự phương pháp MF chuẩn (phần 3.4). Phương pháp này có thể được chính thức hóa:

$$G^0 \approx \hat{G}^0 = W^0 \cdot H^0 \quad (4.1)$$

Để triển khai mô hình học sâu, ta trừ $\hat{G}^0$ cho ma trận ban đầu X , để thu được ma trận mới G^1 có chứa lỗi dự đoán ở lần lặp đầu tiên:

$$G^1 = R - \hat{G}^0 = G^0 - W^0 \cdot H^0 \quad (4.2)$$

Với các giá trị dương trong ma trận G^1 có nghĩa là dự đoán thấp và cần được tăng lên. Tương tự, các giá trị âm trong ma trận G^1 có nghĩa là dự đoán cao và cần được giảm xuống. Sự điều chỉnh này là ý tưởng chính của việc áp dụng phương pháp học sâu. Tiếp tục phân rã ma trận lỗi G^1 như sau:

$$G^1 \approx \hat{G}^1 = W^1 \cdot H^1 \quad (4.3)$$

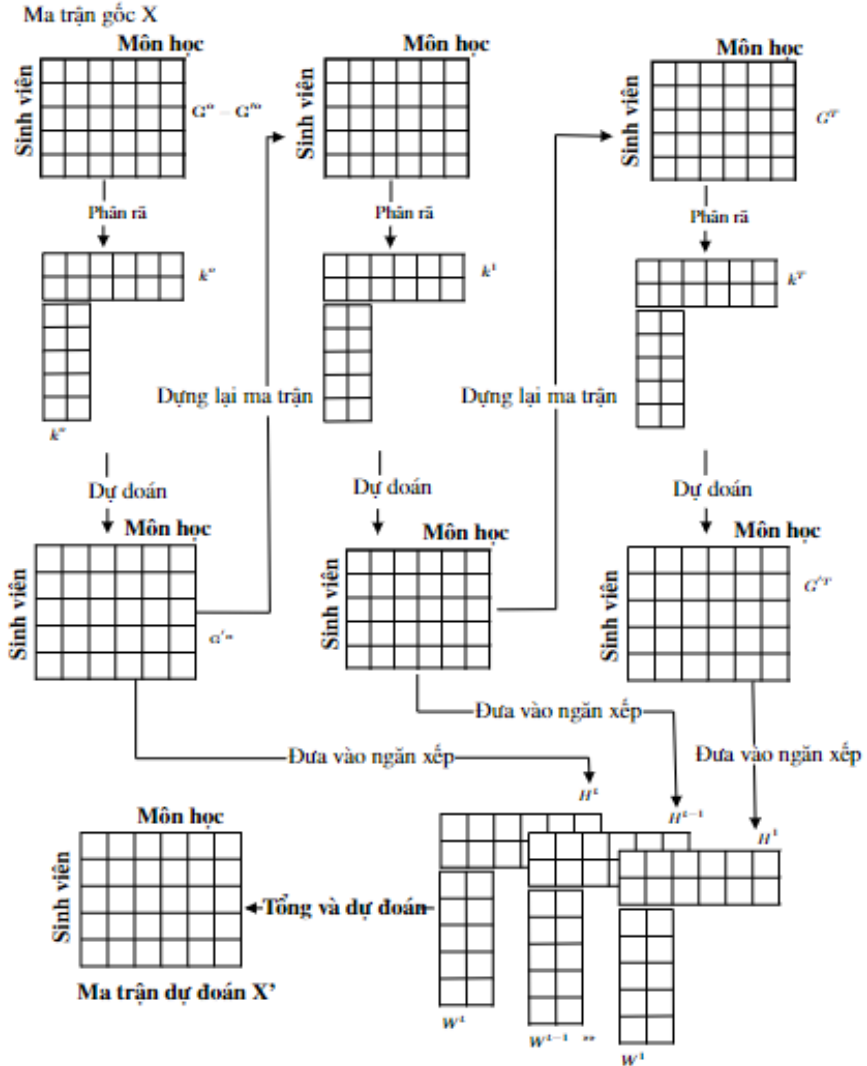
Quá trình xấp xỉ trong mỗi bước được thực hiện tương tự. Tuy nhiên, chúng ta nên lấy $k^1 \neq k^0$ để có được các độ phân giải khác nhau trong quá

trình phân tích nhân tử. Trong trường hợp tổng quát, nếu chúng ta tính toán ở bước $t - 1$ của quy trình học sâu, ma trận lỗi thứ t có thể được xác định:

$$G^t = G^{t-1} - \hat{G}^{t-1} = G^{t-1} - W^{t-1} \cdot H^{t-1} \quad (4.4)$$

Ở bước t , chúng ta cũng phân tích nhân tử thành ma trận W^t và H^t có k^t thừa số tiềm ẩn cho đến khi chuỗi ma trận lỗi hội tụ về không.

$$G^t \approx \hat{G}^t = W^t \cdot H^t \quad (4.5)$$



Hình 4.1: Mô hình phân rã ma trận theo mô hình học sâu

Khi quá trình phân rã sâu kết thúc sau T bước, ma trận phân loại ban đầu X có thể được xây dựng lại bằng cách thêm các ước lượng của sai số:

$$R \approx \hat{R} = \hat{G}^0 + G^1 = \hat{G}^0 + \hat{G}^1 + G^2 = \dots = \hat{G}^0 + \hat{G}^1 + \hat{G}^2 + \dots + \hat{G}^T = W^0 \cdot H^0 + W^1 \cdot H^1 + \dots + W^T \cdot H^T = \sum_{t=0}^T W^t \cdot H^t \quad (4.7)$$

Hàm lỗi cho DMF bây giờ trở thành:

$$O^{DMF} = \sum_{t=0}^T \left(\sum_{(s,c,g) \in D^{train}} (g_{sc}^t - \sum_{k=1}^K w_{sk}^t h_{kc}^t)^2 + \lambda^t (\|W^t\|_F^2 + \|H^t\|_F^2) \right) \quad (4.8)$$

Trong đó thuật ngữ λ^t là siêu tham số chính tắc bước t để tránh việc học vẹt.

Hàm lỗi O^{DMF} có thể được dẫn xuất thành w_{sk}^t và h_{ck}^t dẫn đến các quy tắc được cập nhật sau đây để huấn luyện các tham số mô hình. w_{sk}^t và h_{ck}^t được cập nhật bằng các phương trình bên dưới (trong đó $e_{sc}^t = g_{sc}^t - \hat{g}_{sc}^t$ và $w_{sk}^{t'}$ là giá trị cập nhật của w_{sk}^t và $h_{ck}^{t'}$ là giá trị cập nhật của h_{ck}^t). Với chức năng lỗi mới của DMF, các giá trị của $w_{sk}^{t'}$ và $h_{ck}^{t'}$ được cập nhật tương ứng:

$$h_{ck}^{t'} = h_{ck}^t + \beta(2e_{sc}w_{sk}^t - \lambda h_{ck}^t) \quad (4.9)$$

$$w_{sk}^{t'} = w_{sk}^t + \beta(2e_{sc}h_{ck}^t - \lambda w_{sk}^t) \quad (4.10)$$

Trong đó β^t là tham số tốc độ học của bước t để điều khiển tốc độ học.

Theo cách này, sau khi kết thúc quá trình phân tích nhân tử lồng nhau, tất cả các xếp hạng dự đoán được thu thập trong ma trận $\hat{X} = (\hat{g}_{sc})$, trong đó điểm dự đoán của sinh viên s cho khóa học c được đưa ra bởi công thức sau:

$$\hat{g}_{sc} = \sum_{t=0}^T \sum_{k=1}^K w_{sk}^t h_{kc}^t = \sum_t w_s^t * (h_c^t)^T \quad (4.11)$$

Thuật toán đề xuất

Thuật toán 4.1: Predicting Student Performance DMF

Input: D^{train} , K, betas, lamdas, stopping condition, depth

Output: W, H

1. Let $s \in S$ be a student, $c \in C$ a course, $g \in G$ a grade
2. Let $W[|S|][K], H[|C|][K]$ be latent factors of students, courses
3. $W \leftarrow N(0, \sigma^2)$ and $H \leftarrow N(0, \sigma^2)$
4. $k, t, \beta, \lambda \leftarrow \text{pop}(K, T, \text{Betas}, \text{Lamdas})$

5. while (Stopping criterion is NOT met) do
 6. for each (s, c, g_{sc}) from D^{train}
 7. $\widehat{g}_{sc} \leftarrow \sum_k^K (W[s][k] * H[c][k])$
 8. $e_{sc} = g_{sc} - \widehat{g}_{sc}$
 9. for $k = 1..K$ do
 10. $W[s][k] = W[s][k] + \beta * (2e_{si} * H[c][k] - \lambda * W[s][k])$
 11. $H[c][k] \leftarrow H[c][k] + \beta * (2e_{sc} * W[s][k] - \lambda * H[c][k])$
 12. end for
 13. end for
 14. end while
 15. if (is-empty(K)) return new stack ($\langle W, H \rangle$)
 16. else
 17. $\hat{G} = G - W \cdot H$
 18. Params-Stack=Deep- Matrix-Factoriztion ($\hat{G}, K, T, \text{Betas}, \text{Lamdas}$)
 19. Return push($\langle W, H \rangle$, Params-Stack)
 20. end else.
 21. end function.
-

4.3 Phương pháp phân rã sâu ma trận thiên vị

Tương tự như phương pháp DMF (trình bày ở phần 4.2), phương pháp đề xuất này sử dụng phương pháp Biased-MF làm đầu vào cho quá trình phân rã nhiều tầng thay vì sử dụng MF đơn thuần. Phương pháp này được gọi là phân rã sâu ma trận thiên vị (Deep Biased Matrix Factorization - DBMF). Phương pháp này là tổng hợp kết quả nghiên cứu từ công trình [CT5].

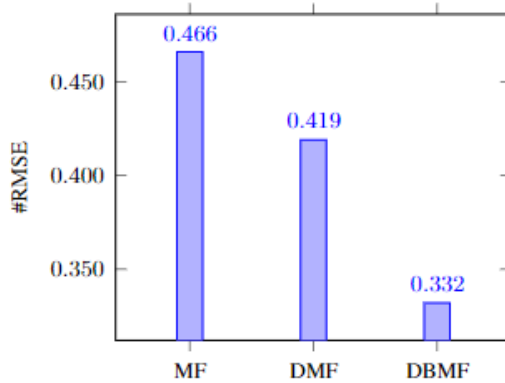
4.4 Đánh giá kết quả

Kết quả thực nghiệm của các phương pháp so sánh được biểu diễn qua (Hình 4.2). Các tham số được dùng cho các phương pháp cũng như độ phức tạp của từng giải thuật tương ứng được trình bày trong bảng 4.1 bên dưới.

Bảng 4.1: Tham số thực nghiệm trên tập dữ liệu ASSISTments

Methods	Tham số	Độ phức tạp
MF	$\beta=0.03$, #iter=50, $K=4$, $\lambda=0.05$	$O(s \times c \times i)$
DMF	Deep (d)= 4; numFactors (k)= [3, 6, 3, 3]; numIters (i)=[50, 50, 50, 50]; learningRate (β)=[0.01, 0.1, 0.1, 0.1]; regularization (λ) = [0.01, 0.1, 0.1, 0.01]	$O(s \times c \times i \times d)$
DBMF	Tham số của DMF, RegS, RegC	$O(s \times c \times i \times d)$

Phương pháp DBMF so sánh với hai phương pháp MF và DMF, được thực nghiệm trên tập dữ liệu Assistment. Biểu đồ cho thấy kết quả RMSE của phương pháp đề xuất (DBMF) là nhỏ nhất (0,332). Sai số nhỏ nhất chứng tỏ mô hình tốt nhất.



Hình 4.2: Kết quả thử nghiệm trên tập dữ liệu ASSISTment

4.5 Tổng kết chương

Chương này đã giới thiệu cách tiếp cận sử dụng kiến trúc học sâu vào phương pháp MF, cũng như Biased-MF nhằm nâng cao độ chính xác dự đoán. Việc áp dụng kiến trúc học sâu để phân rã ma trận khiến thời gian đào tạo sẽ bị chậm lại, nhưng có thể áp dụng thuật toán song song để giải quyết vấn đề này.

Tuy nhiên, các thông tin bổ sung (meta-data) chưa được tận dụng nhằm nâng cao hiệu quả của dự đoán được trình bày ở Chương 5 tiếp theo.

CHƯƠNG 5: TÍCH HỢP CÁC MỐI QUAN HỆ DỮ LIỆU TRONG DỰ ĐOÁN KẾT QUẢ HỌC TẬP

Chương này trình bày chi tiết các phương pháp nâng cao độ chính xác của dự đoán như: tích hợp mối quan hệ của sinh viên (dựa trên mối quan hệ bạn bè cùng lớp học) và mối liên quan của các môn học (dựa trên kiến thức, kỹ năng của các môn học). Kết quả được thực nghiệm trên các tập dữ liệu công bố và nội bộ, đều cho kết quả tốt hơn khi chưa tích hợp [CT2][CT3].

5.1 Đặt vấn đề và các phương pháp giải quyết

Tích hợp mối quan hệ người dùng

Với sự phát triển của mạng xã hội trực tuyến ngày càng tăng làm cho việc tích hợp mạng xã hội vào RS ngày càng được nhiều nhà nghiên cứu quan tâm [20]. Sử dụng việc tích hợp mạng xã hội không giống việc tìm người dùng giống nhau (similar user) mà nó sử dụng thành phần quan hệ của một hệ thống khác như (facebook, twister, zalo...) để bổ sung thông tin, làm cho việc dự đoán đạt kết quả tốt hơn. Với giả định rằng “Nếu A là bạn của B, và A học tốt cũng như chăm học thì có thể ảnh hưởng đến B và ngược lại”. Để thực nghiệm giả định này, luận án đã tích hợp mối quan hệ bạn bè của sinh viên vào mô hình dự đoán năng lực học tập của sinh viên, và đã cho kết quả tốt hơn.

Tích hợp mối liên quan giữa các môn học

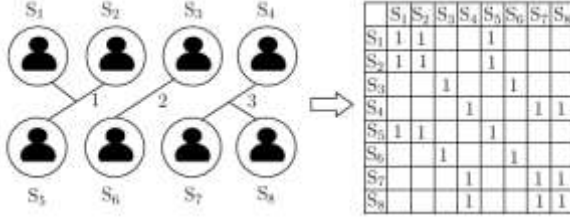
Tương tự như vậy, việc bổ sung mối quan hệ các đối tượng vào mô dự đoán trong hệ thống gợi ý cũng đem lại nhiều hiệu quả dự đoán [21]. Với ý tưởng tương tự: “Nếu hai vấn đề A và B có yêu cầu kỹ năng giống nhau, và một sinh viên giải quyết được vấn đề A thì có thể giải quyết được vấn đề B”. Vì thế, nếu khai thác tốt mối liên quan này có thể nâng cao hiệu quả của bài toán dự đoán năng lực học tập của sinh viên.

Lợi thế về mặt dữ liệu phù hợp với việc tích hợp

Trong tập dữ liệu ASSISTments có trường Student_Class_Id (lớp của sinh viên) dùng làm mối quan hệ bạn bè của sinh viên; và trường Skill_Id (kỹ năng cần đạt của sinh viên) dùng để tìm mối liên quan giữa các môn học. Hai trường này rất có ý nghĩa trong việc nâng cao hiệu quả cho các giải thuật dự đoán.

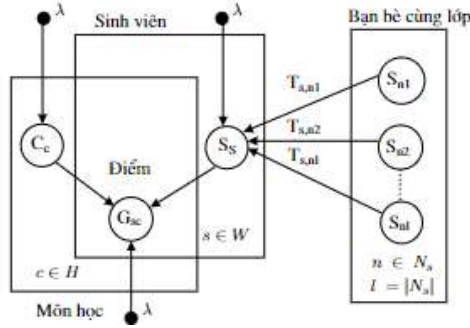
5.2 Phương pháp tích hợp mối quan hệ của sinh viên

Để tích hợp mối quan hệ của người học vào trong kỹ thuật phân rã ma trận, mối quan hệ bạn bè của các sinh viên trong cùng một lớp sẽ được chuyển thành ma trận quan hệ dạng nhị phân (xem Hình 5.1).



Hình 5.1: Sự chuyển đổi mối quan hệ bạn bè thành ma trận nhị phân

Trong tiếp cận tích hợp mối quan hệ, mỗi sinh viên s có một tập hợp N_s thể hiện các mối quan hệ bạn bè với s . Giá trị $T_{s,n}$ tỷ lệ xác suất thể hiện mối quan hệ của sinh viên s với sinh viên n , có giá trị là một số nhị phân, nếu $T = 0$ thì hai người s và n khác lớp, nếu $T = 1$ thì hai người cùng lớp. Hình 5.2 thể hiện mô hình tích hợp mối quan hệ người dùng vào kỹ thuật MF.



Hình 5.2: Mô hình tích hợp mối quan hệ người dùng vào kỹ thuật MF

Để tích hợp được mối quan hệ, chúng ta chỉ cần thay thế ma trận W bằng $\hat{W}$ bằng công thức (5.1) và sau đó thực hiện tương tự kỹ thuật MF cơ sở.

$$\hat{W}_s = \frac{\sum_{n \in N_s} T_{s,n} w_n}{\sum_{n \in N_s} T_{s,n}} = \frac{1}{|N_s|} \sum_{n \in N_s} w_n \quad (5.1)$$

Do ma trận T là ma trận nhị phân nên mỗi phần tử $T_{s,n}$ có giá trị là 1 hoặc 0, nên $\sum_{n \in N_s} T_{s,n} = |N_s|$ và có thể hiểu đơn giản là $\hat{W}$ bằng với giá trị

trung bình của các người dùng có quan hệ. Việc dự đoán vẫn áp dụng theo giải thuật MF nhưng thay thế W bằng $\hat{W}$. Như vậy, hàm dự đoán trở thành:

$$\hat{g}_{sc} = \sum_{k=1}^K \hat{w}_{sk} h_{ck} = \hat{w}_s \cdot h_c^T \quad (5.2)$$

Hàm mục tiêu của kỹ thuật SocialMF vẫn được tính như kỹ thuật MF nhưng có cộng thêm một hệ số của mối quan hệ:

$$O^{\text{SocialMF}} = \sum_{(s,c,g) \in D^{\text{train}}} (g_{sc} - \sum_{k=1}^K \hat{w}_{sk} h_{ck})^2 + \lambda (\|W\|_F^2 + \|H\|_F^2) + \lambda_T \sum_{s=1}^S \left(w_s - \frac{1}{|N_s|} \sum_{n \in N_s} w_n \right)^2 \quad (5.3)$$

Các giá trị w và h cũng được cập nhật mới sau mỗi bước lặp như sau:

$$w'_{sk} = w_{sk} + \beta (2e_{sc} h_{ck} - \lambda w_{sk}) + \lambda_T \left(w_{sk} - \frac{1}{|N_s|} \sum_{n \in N_s} w_{nk} \right) - \lambda_T \left(\sum_{t \in S_s} \frac{1}{|N_s|} \left(w_{tk} - \frac{1}{|N_s|} \sum_{w \in N_s} w_{wk} \right) \right) \quad (5.4)$$

$$h'_{ck} = h_{ck} + \beta (2e_{sc} w_{sk} - \lambda h_{ck}) \quad (5.5)$$

Với $e_{sc} = g_{sc} - \hat{g}_{sc}$, β là tốc độ học, λ là hệ số chính tắc cho ma trận W và H , λ_T là hệ số chính tắc cho ma trận mối quan hệ T , tập S_s là tập tất cả sinh viên, N_s là tập các sinh viên cùng lớp, $|N_s|$ là tổng số sinh viên cùng lớp với sinh viên s . Sau quá trình tối ưu, ta nhận được các giá trị của ma trận W và H đã tối ưu. Khi đó, có thể dễ dàng dự đoán kết quả thông qua công thức:

$$\hat{g}_{sc} = \sum_{k=1}^K w_{sk} h_{ck} = w_s \cdot h_c^T \quad (5.6)$$

5.3 Phương pháp tích hợp mối liên quan giữa các môn học

Tương tự, ta cần chuyển mối liên quan này về ma trận quan hệ nhị phân R . Mỗi môn học có tập hợp R_c thể hiện mối liên quan giữa môn học c với môn học m . Giá trị R_{cm} có giá trị 0 hoặc 1. Nếu $R_{cm} = 1$ thì hai môn học c và m có liên quan nhau, còn $R_{cm} = 0$ là ngược lại. Mô hình MF vẫn làm nền tảng cho việc tích hợp mối liên quan giữa các môn học vào mô hình.

Tương tự như SocialMF, ta tính giá trị $\hat{h}$ thay thế cho h với công thức:

$$\hat{h}_c = \frac{\sum_{m \in R_c} R_{cm} h_m}{\sum_{m \in R_c} R_{cm}} = \frac{1}{|R_c|} \sum_{m \in R_c} h_m \quad (5.7)$$

Việc dự đoán vẫn áp dụng theo công thức của giải thuật MF nhưng thay thế H bằng $\hat{H}$. Như vậy, công thức dự đoán trở thành:

$$\hat{g}_{sc} = \sum_{k=1}^K w_{sk} \hat{h}_{ck} = w_s \cdot \hat{h}_c^T \quad (5.8)$$

Hàm mục tiêu sau khi tích hợp mối quan hệ giữa các môn học được tính theo biểu thức như sau:

$$O^{\text{Item-MF}} = \sum_{(s,c,g) \in D^{\text{train}}} (g_{sc} - \sum_{k=1}^K \hat{w}_{sk} \hat{h}_{ck})^2 + \lambda (\|W\|_F^2 + \|H\|_F^2) + \lambda_R \sum_{c=1}^C \left(h_c - \frac{1}{|M_c|} \sum_{m \in M_c} h_m \right)^2 \quad (5.9)$$

Các giá trị w và h cũng được cập nhật mới sau mỗi bước lặp:

$$w'_{sk} = w_{sk} + \beta (2e_{sc} h_{ck} - \lambda w_{sk}) \quad (5.10)$$

$$h'_{ck} = h_{ck} + \beta (2e_{sc} w_{sk} - \lambda h_{ck}) + \lambda_R \left(h_{ck} - \frac{1}{|M_c|} \sum_{m \in M_c} h_{mk} \right) - \lambda_R \left(\sum_{r \in C_c} \frac{1}{|M_c|} \left(h_{rk} - \frac{1}{|M_c|} \sum_{y \in R_c} h_{yk} \right) \right) \quad (5.11)$$

Với $e_{sc} = g_{sc} - \hat{g}_{sc}$, $\hat{a}$ là tốc độ học, ϵ là hệ số chính tắc cho ma trận W và H , λ_R là hệ số chính tắc cho ma trận liên quan R , tập C_c là tập tất cả môn học, M_c là tập các môn học có liên quan với nhau, $|M_c|$ là tổng số các môn học cùng nhóm liên quan. Sau quá trình tối ưu, công thức dự đoán kết quả như sau:

$$\hat{g}_{sc} = w_s \cdot \hat{h}_c^T \quad (5.12)$$

5.4 Đánh giá kết quả

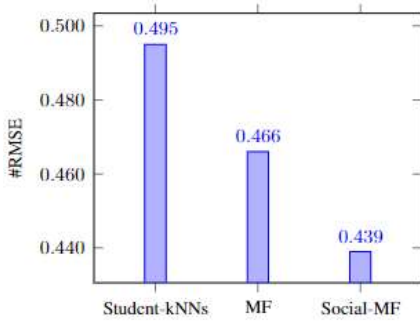
Tập dữ liệu ASSISTments sau khi xử lý bao gồm 8519 sinh viên, 35978 bài tập, 1011079 đánh giá năng lực. Tập dữ liệu điểm trường Đại học Cần Thơ (CTU), bao gồm 4017 sinh viên, 353 môn học và gồm 279536 điểm chi tiết.

Các siêu tham số của phương pháp và độ phức tạp của từng giải thuật tương ứng đều được mô tả chi tiết trong bảng 5.1 bên dưới.

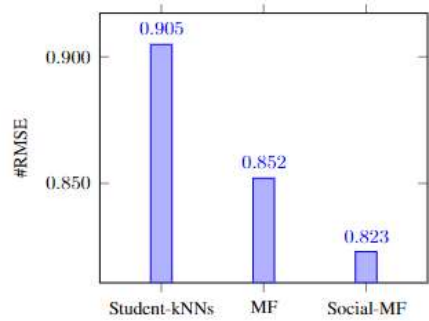
Bảng 5.1: Hyper-parameters on ASSISTments dataset

Methods	Hyper-Parameters	Độ phức tạp
Student-kNNs	k=5, simMeasure=Cosine	$O(s^2 \times c)$
Course-kNNs	k=5, simMeasure=Cosine	$O(c^2 \times s)$
MF	$\beta=0.01$, #iter=60, K=32, $\lambda=0.1$	$O(s \times c \times i)$
Social-MF	K=80, $\lambda=3$, $\beta=0.01$, SocialReg=5, i=500	$O(s^3 \times c \times i)$
CRMF	K=80, $\lambda=3$, $\beta=0.01$, ItemReg=3, i=300	$O(c^3 \times s \times i)$

Từ biểu đồ so sánh (Hình 5.4) cho thấy khi áp dụng giải thuật tích hợp mối quan hệ người dùng vào bài toán dự đoán kết quả học tập của sinh viên được thực nghiệm trên 2 tập dữ liệu ở Việt Nam và Quốc tế đều đạt được độ lỗi RMSE thấp nhất so với các giải thuật khác.



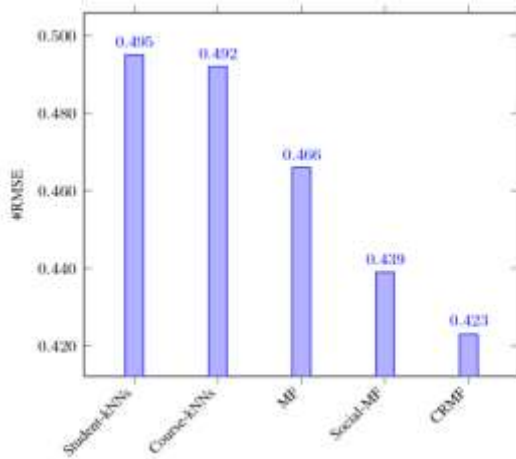
(a) Tập dữ liệu ASSISTment



(b) Tập dữ liệu CTU

Hình 5.4: So sánh độ lỗi RMSE trên 2 tập dữ liệu ở Việt Nam và Quốc tế

Tương tự, Từ biểu đồ (Hình 5.3) cho thấy phương pháp CRMF giải bài toán dự đoán kết quả sinh viên đạt độ lỗi RMSE thấp nhất so với các giải thuật khác (RMSE của CRMF là 0.423).



Hình 5.3: Bảng so sánh độ lỗi RMSE của các giải thuật tích hợp

5.5 Tổng kết chương

Luận án đã giới thiệu hai hướng tiếp cận tích hợp mối quan hệ bạn bè của người học và mối liên quan của các môn học vào dự đoán kết quả học tập sinh viên, được thực hiện qua hai kỹ thuật SocialMF và CRMF. Với các cách tiếp cận này, chúng ta có thể tận dụng được thông tin khác từ tập dữ liệu cũ để tích hợp vào xây dựng mô hình dự đoán, cải thiện được kết quả dự đoán.

CHƯƠNG 6: KẾT LUẬN VÀ KIẾN NGHỊ

6.1 Kết luận

Luận án hướng tới một chủ đề có ý nghĩa về lý thuyết và thực tiễn của khoa học máy tính được cộng đồng nghiên cứu và nhà giáo dục đặc biệt quan tâm, đó là nghiên cứu các giải pháp gợi ý trong cố vấn học tập và xây dựng ứng dụng thử nghiệm.

Đóng góp về mặt lý thuyết

Luận án đã trình bày những nghiên cứu cơ bản và mở rộng về các kỹ thuật khai phá dữ liệu giáo dục trong dự đoán năng lực học tập của sinh viên theo hướng cá nhân hóa, và cho thấy việc ứng dụng hệ thống gợi ý trong cố vấn học tập đã phần nào hỗ trợ tích cực giúp đỡ sinh viên lựa chọn những môn học thích hợp với từng cá nhân sinh viên cụ thể.

Mặc dù ứng dụng hệ thống gợi ý trong dự đoán kết quả học tập của sinh viên có kết quả khả quan, tuy nhiên vấn đề về nâng cao độ chính xác của giải thuật cũng như tận dụng thông tin bổ sung thì vẫn còn bỏ ngỏ. Vì vậy, Luận án đã đề xuất một số phương pháp tích hợp mối quan hệ dữ liệu vào kỹ thuật phân rã ma trận nhằm nâng cao hiệu quả trong dự đoán. Thật vậy, việc tích hợp các thông tin tương tác (mối quan hệ bạn bè của sinh viên, mối liên quan giữa các môn học) vào mô hình dự đoán làm tăng thêm độ chính xác của dự đoán. Qua những thí nghiệm cho thấy các phương pháp tích hợp mang lại hiệu quả cao hơn các phương pháp khác.

Hiện nay kỹ thuật học sâu là kỹ thuật phổ biến trong các bài toán dự đoán, nhận dạng, phân loại được áp dụng trong nhiều lĩnh vực. Với ý tưởng dựa trên sự ưu việt của kỹ thuật học sâu, luận án kết hợp kiến trúc học sâu với kỹ thuật phân rã ma trận trong hệ thống gợi ý, nhằm đưa ra kết quả dự đoán chính xác hơn. Từ đó có thể áp dụng vào bài toán dự đoán kết quả học tập của sinh viên để làm cơ sở gợi ý lựa chọn môn học phù hợp.

Đóng góp về mặt thực tiễn

Về mặt thực tiễn, để ứng dụng được các kỹ thuật khai phá dữ liệu vào giải bài toán dự đoán năng lực sinh viên và hỗ trợ cố vấn học tập gợi ý thì cần xây dựng một công cụ trực quan (website) hỗ trợ cho sinh viên ra quyết định

lập kế hoạch học tập phù hợp với khả năng của riêng mình. Chính vì thế luận án đã đề xuất kiến trúc mô hình để xây dựng hệ thống gợi ý lựa chọn môn học một cách linh động, dễ dàng nâng cấp hoặc cải tiến.

6.2 Hạn chế và hướng phát triển

Hạn chế

Mặc dù các đề xuất được đưa ra bởi luận án đã giải quyết được bài toán gợi ý sinh viên lựa chọn môn học, một vấn đề quan trọng và tốn nhiều công sức của cố vấn học tập. Tuy nhiên giải pháp còn tồn tại một số hạn chế nhất định chưa được giải quyết trong các đề xuất được đưa ra.

- Vấn đề gợi ý cho sinh viên lựa chọn các môn học theo thứ tự phù hợp với từng sinh viên..
- Vấn đề dữ liệu bổ sung còn hạn chế như: dữ liệu mạng xã hội của sinh viên, hoặc dữ liệu số hóa nội dung của các môn học.

Hướng phát triển

Năng lực học tập của sinh viên sẽ thay đổi theo thời gian làm ảnh hưởng đến kết quả dự đoán. Do đó việc nghiên cứu kỹ thuật phân rã ma trận theo thời gian (Time Series MF) sẽ là một hướng phát triển của nghiên cứu.

Ngoài ra, thứ tự của các môn học mà sinh viên học sẽ ảnh hưởng đến kết quả học tập của sinh viên cũng như kết quả dự đoán của giải thuật. Việc dự đoán kết quả theo kỹ thuật phân rã ma trận tuần tự (Sequence MF) cũng là một trong những hướng phát triển đầy tiềm năng.

Bên cạnh việc ứng dụng kỹ thuật phân rã ma trận trong dự đoán kết quả học tập của sinh viên đạt được nhiều thành công, thì việc ứng dụng các phương pháp hiện đại khác để giải quyết bài toán này còn bỏ ngỏ như: độ đo khoảng cách trong hệ thống gợi ý; hoặc phương pháp mới hiện nay như mạng thần kinh đồ thị (Graph Neural Networks) áp dụng sức mạnh dự đoán của học sâu cho các cấu trúc dữ liệu phong phú mô tả những đối tượng và mối quan hệ của chúng dưới dạng các điểm được kết nối bằng các đường trong biểu đồ. Trong tương lai, đây là những phương pháp tiềm năng để giải quyết bài toán dự đoán kết quả học tập của sinh viên.

TÀI LIỆU THAM KHẢO

- [1] A. C. Rivera, M. Tapia-Leon, and S. Lujan-Mora, "Recommendation Systems in Education: A Systematic Mapping Study," *Advances in Intelligent Systems and Computing*, vol. 721, pp. 937–947, jan 2018. https://link.springer.com/chapter/10.1007/978-3-319-73450-7_89
- [2] J. S. Gil, A. J. P. Delima, and R. N. Vilchez, "Predicting students' dropout indicators in public school using data mining approaches," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 1, pp. 774–778, 2020, doi: 10.30534/ijatcse/2020/110912020.
- [3] J. Kabathova and M. Drlik, "Towards Predicting Student's Dropout in University Courses Using Different Machine Learning Techniques," *Appl. Sci. 2021, Vol. 11, Page 3130*, vol. 11, no. 7, p. 3130, Apr. 2021, doi: 10.3390/APP11073130.
- [4] H. Q. Nguyen, T. T. Pham, V. Vo, B. Vo, and T. T. Quan, "The predictive modeling for learning student results based on sequential rules," *Int. J. Innov. Comput. Inf. Control*, vol. 14, no. 6, pp. 2129–2140, Dec. 2018, doi: 10.24507/IJICIC.14.06.2129.
- [5] H. H. Varzaneh, B. S. Neysiani, H. Ziafat, and N. Soltani, "Recommendation Systems Based on Association Rule Mining for a Target Object by Evolutionary Algorithms," *Emerg. Sci. J.*, vol. 2, no. 2, pp. 100–107, May 2018, doi: 10.28991/ESJ-2018-01133.
- [6] A. Esteban, A. Zafra, and C. Romero, "Helping university students to choose elective courses by using a hybrid multi-criteria recommendation system with genetic optimization," *Knowledge-Based Syst.*, vol. 194, p. 105385, Apr. 2020, doi: 10.1016/J.KNOSYS.2019.105385.
- [7] B. C. Madeira, T. Tasci, and N. Celebi, "Prediction of Student Performance Using Rough Set Theory And Backpropagation Neural Networks," *Eur. Sci. Journal, ESJ*, vol. 17, no. 7, pp. 1–1, Feb. 2021, doi: 10.19044/ESJ.2021.V17N7P1.
- [8] T. T. Dien, S. H. Luu, N. Thanh-Hai, and N. Thai-Nghe, "Deep Learning with Data Transformation and Factor Analysis for Student Performance Prediction," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 8, pp. 711–721, 2020, doi: 10.14569/IJACSA.2020.0110886.
- [9] X. Li, X. Zhu, X. Zhu, Y. Ji, and X. Tang, "Student Academic Performance Prediction Using Deep Multi-source Behavior Sequential Network," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12084 LNAI, pp. 567–579, May 2020, doi: 10.1007/978-3-030-47426-3_44.
- [10] M. Hasnawi, N. Kurniati, S. H. Mansyur, Irawati, and T. Hasanuddin, "Combination of Case Based Reasoning with Nearest Neighbor and Decision Tree for Early Warning System of Student Achievement," *Proc. - 2nd East Indones. Conf. Comput. Inf. Technol. Internet Things*

Ind. EIconCIT 2018, pp. 78–81, Nov. 2018, doi: 10.1109/EICONCIT.2018.8878512.

- [11] N. Thai-Nghe and L. Schmidt-Thieme, “Multi-relational Factorization Models for Student Modeling in Intelligent Tutoring Systems,” in *Proceedings - 2015 IEEE International Conference on Knowledge and Systems Engineering, KSE 2015*, Jan. 2015, pp. 61–66, doi: 10.1109/KSE.2015.9.
- [12] B. Park Woolf, L. New York, S. Diego, and M. Kaufmann, *Building Intelligent Interactive Tutors Student-centered strategies for revolutionizing e-learning*. 2009.
- [13] X. Du, J. Yang, J. L. Hung, and B. Shelton, “Educational data mining: a systematic review of research and emerging trends,” *Inf. Discov. Deliv.*, vol. 48, no. 4, pp. 225–236, Oct. 2020, doi: 10.1108/IDD-09-2019-0070/FULL/XML.
- [14] M. C. Mihaescu and P. S. Popescu, “Review on publicly available datasets for educational data mining,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 11, no. 3, p. e1403, May 2021, doi: 10.1002/WIDM.1403.
- [15] N. Thai-Nghe, L. Drumond, A. Krohn-Grimberghe, and L. Schmidt-Thieme, “Recommender system for predicting student performance,” *Procedia Comput. Sci.*, vol. 1, no. 2, pp. 2811–2819, 2010, doi: 10.1016/J.PROCS.2010.08.006.
- [16] R. Lara-Cabrera, Á. González-Prieto, and F. Ortega, “Deep Matrix Factorization Approach for Collaborative Filtering Recommender Systems,” *Appl. Sci. 2020, Vol. 10, Page 4926*, vol. 10, no. 14, p. 4926, Jul. 2020, doi: 10.3390/APP10144926.
- [17] T. T. Dien, N. Thanh-Hai, and N. T. Nghe, “Deep Matrix Factorization for Learning Resources Recommendation,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12876 LNAI, pp. 167–179, Sep. 2021, doi: 10.1007/978-3-030-88081-1_13.
- [18] X. Su and T. M. Khoshgoftaar, “A Survey of Collaborative Filtering Techniques,” *Advances in Artificial Intelligence*, vol. 2009, pp. 1–19.
- [19] Y. Koren, R. Bell, and C. Volinsky, “Matrix factorization techniques for recommender systems,” *Computer (Long. Beach. Calif.)*, vol. 42, no. 8, pp. 30–37, 2009, doi: 10.1109/MC.2009.263.
- [20] R. Chen *et al.*, “A Novel Social Recommendation Method Fusing User’s Social Status and Homophily Based on Matrix Factorization Techniques,” *IEEE Access*, vol. 7, pp. 18783–18798, 2019, doi: 10.1109/ACCESS.2019.2893024.
- [21] Y. Yu, C. Wang, H. Wang, and Y. Gao, “Attributes coupling based matrix factorization for item recommendation,” *Appl. Intell.*, vol. 46, no. 3, pp. 521–533, Apr. 2017, doi: 10.1007/S10489-016-0841-8.

DANH MỤC CÔNG TRÌNH ĐÃ CÔNG BỐ

[CT1] Huynh-Ly Thanh-Nhan, Huu-Hoa Nguyen, and Nguyen Thai-Nghe. (2016). Methods for building course recommendation systems. In Proceedings of the 2016 International Conference on Knowledge and Systems Engineering (KSE 2016, ISBN 978-1-4673-8929-7, **IEEE Xplore. Scopus indexed.**

[CT2] Huynh-Ly Thanh-Nhan, Le Huy-Thap, and Nguyen Thai-Nghe. (2017). Toward integrating social networks into Intelligent Tutoring Systems. In Proceedings of the 2017 International Conference on Knowledge and Systems Engineering, (KSE 2017), ISBN 978-1-5386-3576-6, **IEEE Xplore. Scopus indexed.**

[CT3] Huynh-Ly Thanh-Nhan, Le Huy-Thap and Nguyen Thai-Nghe. (2020). Integrating courses' relationship into predicting student performance, International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), pp. 6375-6383, Vol. 9, No 4, 2020. ISSN: 2278–3091. **Scopus indexed 2020 .**

[CT4] Huynh-Ly TN., Le HT., Thai-Nghe N. (2021). Integrating Deep Learning Architecture into Matrix Factorization for Student Performance Prediction. In: Dang T.K., Küng J., Chung T.M., Takizawa M. (eds) Future Data and Security Engineering. FDSE 2021. Lecture Notes in Computer Science, vol 13076. **Springer, Cham. Scopus Q2**

[CT5] T.-N. Huynh-Ly, H.-T. Le, and N. Thai-Nghe, (2023), Deep Biased Matrix Factorization for Student Performance Prediction, EAI Endorsed Trans Context Aware Syst App, vol. 9, no. 1, Apr. 2023.