

**BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC LẠC HỒNG**



HUỲNH LÝ THANH NHÀN

**GIẢI PHÁP GỢI Ý TRONG VIỆC
CỔ VẤN HỌC TẬP**

LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

Đồng Nai, năm 2024

**BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC LẠC HỒNG**



HUỲNH LÝ THANH NHÀN

**GIẢI PHÁP GỢI Ý TRONG VIỆC
CỔ VẤN HỌC TẬP**

LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

Bản luận án được bảo vệ cấp trường

Chuyên ngành: Khoa học Máy tính

Mã số ngành: 9480101

NGƯỜI HƯỚNG DẪN KHOA HỌC

1. PGS. TS. NGUYỄN THÁI NGHE
2. PGS. TS. LÊ HUY THẬP

Đồng Nai, năm 2024

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu do tôi thực hiện dưới sự hướng dẫn của PGS. TS. Nguyễn Thái Nghe và PGS. TS. Lê Huy Thập. Tôi cam đoan các kết quả nghiên cứu được trình bày trong luận án là trung thực và không sao chép từ bất kỳ công trình nghiên cứu nào khác. Một số kết quả là thành quả tập thể và đã được các đồng tác giả đồng ý cho sử dụng trong luận án. Mọi trích dẫn trong luận án đều có ghi nguồn gốc xuất xứ rõ ràng và đầy đủ.

Đồng Nai, tháng 01 năm 2024

Người viết

Huỳnh Lý Thanh Nhân

LỜI CẢM ƠN

Để hoàn thành luận án này, tôi đã nhận được sự hướng dẫn tận tình, quan tâm đặc biệt, giúp đỡ nhiệt tình từ quý Thầy Cô, đồng nghiệp, bạn bè và người thân. Tôi xin gửi lời cảm ơn chân thành và sâu sắc đến:

Thầy PGS.TS. Nguyễn Thái Nghe và Thầy PGS.TS. Lê Huy Thập đã tận tình chỉ bảo, hướng dẫn, chỉnh sửa, động viên và tạo mọi điều kiện tốt nhất cho tôi trong quá trình học tập và nghiên cứu.

Ban giám hiệu trường Đại học Lạc Hồng, quý Thầy, Cô và anh chị của Khoa CNTT và Khoa Sau Đại học của Trường Đại học Lạc Hồng, đã cung cấp kiến thức, hướng dẫn, và tạo mọi điều kiện cho tôi và cũng như đặc biệt quan tâm và hỗ trợ tôi trong suốt quá trình học tập.

Ban giám hiệu trường Đại học An Giang Đại học Quốc gia TP. HCM, Ban chủ nhiệm khoa Công nghệ Thông tin, Bộ môn Kỹ thuật phần mềm đã tạo điều kiện để tôi tham gia học tập nâng cao trình độ chuyên môn, các bạn đồng nghiệp đã không ngừng động viên và giúp đỡ tôi trong suốt thời gian học tập.

Sau cùng tôi xin chân thành cảm ơn sâu sắc đến gia đình và người thân đã giúp đỡ và luôn động viên tôi trong suốt quá trình học tập và tạo điều kiện tốt nhất để tôi hoàn thành luận án.

NCS Huỳnh Lý Thanh Nhân

TÓM TẮT

Gần đây, ở các trường Cao đẳng Đại học, số lượng sinh viên bị cảnh báo học vụ và buộc thôi học có chiều hướng gia tăng. Đây là một tổn thất lớn cho sinh viên, gia đình, nhà trường và xã hội. Một trong những nguyên nhân hàng đầu là do sinh viên không tự đoán trước được năng lực của mình cũng như lựa chọn môn học không hợp lý để có kế hoạch học tập phù hợp theo khả năng của họ. Nhằm đưa ra những giải pháp gợi ý lựa chọn môn học tự chọn trong việc cố vấn học tập là vấn đề cấp thiết cần được giải quyết. Bên cạnh đó, nhờ vào sự phát triển của trí tuệ nhân tạo nói chung và hệ thống gợi ý nói riêng đã giải quyết nhiều vấn đề thực tiễn trong gợi ý học tập của cố vấn học tập. Tuy nhiên, kết quả của những nghiên cứu này chưa cao vì chưa khai thác hết các mối quan hệ dữ liệu và chưa cải tiến mô hình dự đoán. Đây cũng là lý do để chúng tôi thực hiện nghiên cứu. Trong luận án này, chúng tôi đã giải quyết các vấn đề bằng những đóng góp sau:

Thiết kế mô hình dự đoán kết quả học tập của sinh viên theo hướng tiếp cận hệ thống gợi ý. Mỗi sinh viên sẽ được dự đoán kết quả học tập của tất cả các môn học mà sinh viên này chưa học, sau đó sẽ lọc ra những môn học được dự đoán với kết quả cao để gợi ý cho sinh viên lựa chọn. Kết quả dự đoán của phương pháp này là khá tốt và hoàn toàn có thể áp dụng rộng rãi trong việc cố vấn học tập tự động.

Nhằm nâng cao hiệu quả của dự đoán, nhiều nhà nghiên cứu cũng quan tâm nhiều đến nguồn dữ liệu bổ sung. Theo xu hướng này, luận án đã đề xuất một số phương pháp tích hợp như: tích hợp mối quan hệ bạn bè giữa những người học cùng lớp với nhau; tích hợp mối liên quan giữa các môn học nhờ vào chủ đề hay nội dung kiến thức của những môn học liên quan nhau. Việc tích hợp các nguồn dữ liệu bổ sung này, làm cho mô hình học tăng cường và đã đưa ra dự đoán chính xác hơn các phương pháp phân rã ma trận thông thường.

Hiện nay kỹ thuật học sâu đang là những tuyệt tác 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, chúng tôi 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 chọn môn học tự chọn một cách phù hợp.

Qua kết quả thử nghiệm trên các giải thuật khác nhau và các tập dữ liệu khác nhau cho thấy những đề xuất của luận án có kết quả dự đoán chính xác hơn. Ứng dụng thử nghiệm sử dụng các phương pháp đã đề xuất có thể hoàn toàn ứng dụng cho việc gợi ý lựa chọn môn học của cố vấn học tập ở các trường học mong muốn bắt kịp xu hướng của cuộc cách mạng công nghệ lần thứ tư trong giáo dục.

Từ khóa: Dự đoán điểm sinh viên, Khai phá dữ liệu giáo dục, Giải pháp gợi ý cho cố vấn học tập

ABSTRACT

Recently, in colleges and universities, the number of students warned and forced to drop out of school has been on the rise. This problem is a great loss for students, families, schools and society. One of the top reasons is that students can't predict their abilities and choose unreasonable subjects to have an appropriate study plan according to their ability. To provide course recommendation in the academic counselling is an urgent problem that needs to be solved. In addition, thanks to the development of artificial intelligence in general and the suggestion system in particular, many practical issues have been solved in the study advisor's suggestions. However, the results of these studies are not high because they have not fully exploited the data relationships and have not improved the predictive model. This is also the reason for us to do the research. In this thesis, we have solved these problems with the following contributions:

Design a model to predict student learning outcomes in the direction of a suggested system approach. The system will predict the learning results of all subjects that this student has not studied and then will filter out the predicted subjects with high marks to suggest students choose. The prediction results of this method are quite good and can be widely applied in automatic learning mentoring.

In order to improve the efficiency of predictions, many researchers are also interested in additional data sources. Following this trend, the thesis has proposed several integration methods such as: integrating the friendship relationship between students in the same class; integrate the relationship between subjects thanks to the subject or content of subject knowledge. The integration of these additional data sources made the learning model more robust and gave more accurate predictions than the standard matrix factorization method.

Currently, deep learning techniques are the masterpieces in prediction, identification and classification problems applied in many fields. With the idea based on the superiority of the deep learning technique, we combine deep learning architecture with matrix decomposition technique in the recommender system to give more accurate prediction

results. From there, it can be applied to the problem of predicting student learning outcomes to serve as a basis for choosing elective subjects appropriately.

The thesis's proposals have more accurate prediction results by testing results on different algorithms and data sets. The experimental application using the proposed methods can be applied entirely to suggest the subject selection of academic advisors in schools wishing to catch up with the industry 4.0 in education.

Key words: Predicting Student Performance, Educational Data Mining, Academic advising

MỤC LỤC

LỜI CẢM ƠN	ii
TÓM TẮT	iii
ABSTRACT	v
MỤC LỤC	vii
DANH MỤC CÁC HÌNH VẼ	xiii
DANH MỤC CÁC BẢNG BIỂU	xiv
 CHƯƠNG 1 GIỚI THIỆU	 1
1.1 Tính cấp thiết của luận án	1
1.2 Bài toán nghiên cứu	3
1.2.1 Sự tương quan giữa bài toán xếp hạng và bài toán dự đoán kết quả học tập	3
1.2.2 Bài toán dự đoán kết quả học tập của sinh viên	4
1.2.3 Gợi ý lựa chọn môn học tự chọn	5
1.3 Thách thức của bài toán nghiên cứu	7
1.4 Các vấn đề nghiên cứu	8
1.5 Mục tiêu nghiên cứu và hướng tiếp cận của luận án	8
1.6 Đối tượng, phương pháp và phạm vi nghiên cứu	9
1.7 Các đóng góp của luận án	10
1.8 Bố cục của luận án	11
 CHƯƠNG 2 CƠ SỞ LÝ THUYẾT VÀ CÁC NGHIÊN CỨU LIÊN QUAN	 13
2.1 Các nghiên cứu liên quan	13
2.1.1 Các nghiên cứu về khai thác dữ liệu giáo dục	13
2.1.2 Các nghiên cứu về hệ thống gợi ý	15
2.1.3 Các nghiên cứu về dự đoán năng lực học tập của sinh viên	20
2.1.4 Các nghiên cứu về cải tiến các mô hình dự đoán	21
2.1.5 Các nghiên cứu về tích hợp mối quan hệ dữ liệu	22

2.2	Tổng quan về hệ thống gợi ý	23
2.2.1	Giới thiệu về hệ thống gợi ý	23
2.2.2	Các lĩnh vực ứng dụng của hệ thống gợi ý	24
2.2.3	Giới thiệu hệ thống trợ giảng thông minh	25
2.2.4	Các nguồn tài nguyên về hệ thống gợi ý	26
2.2.5	Các hướng tiếp cận để xây dựng hệ thống gợi ý	28
2.3	Bài toán dự đoán kết quả học tập của sinh viên	35
2.3.1	Dự đoán theo cá nhân hóa	35
2.3.2	Dự đoán quy luật chung (không cá nhân hóa)	38
2.3.3	Các giải pháp gợi ý của cổ vấn học tập	40
2.4	Tổng kết chương	41

CHƯƠNG 3 DỰ ĐOÁN KẾT QUẢ HỌC TẬP CỦA SINH VIÊN THEO HƯỚNG TIẾP CẬN HỆ THỐNG GỢI Ý 43

3.1	Giải bài toán dự đoán kết quả học tập sinh viên theo hướng tiếp cận hệ thống gợi ý	43
3.1.1	Phát biểu bài toán hệ thống gợi ý trong ngữ cảnh giáo dục	43
3.1.2	Quy trình xây dựng hệ thống gợi ý	45
3.1.3	Đánh giá hệ thống gợi ý	46
3.1.4	Độ đo đánh giá độ chính xác của dự đoán kết quả học tập	47
3.2	Phương pháp lọc cộng tác dựa vào sinh viên tương tự (Student-kNNs)	48
3.3	Phương pháp lọc cộng tác dựa vào môn học tương tự (Course-kNNs)	52
3.4	Phương pháp phân rã ma trận trong dự đoán kết quả học tập sinh viên (PSP-MF)	55
3.5	Phương pháp phân rã ma trận thiên vị - Biased Matrix Factorization	58
3.6	Đánh giá kết quả	63
3.6.1	Tập dữ liệu	63
3.6.2	Cài đặt thực nghiệm	64
3.6.3	Kết quả thử nghiệm	64
3.7	Tổng kết chương	66

CHƯƠNG 4	CÁC MÔ HÌNH PHÂN RÃ SÂU MA TRẬN ĐỂ DỰ ĐOÁN	
	KẾT QUẢ HỌC TẬP CỦA SINH VIÊN	68
4.1	Bài toán nghiên cứu	68
4.2	Kỹ thuật phân rã sâu ma trận trong bài toán dự đoán kết quả học tập của sinh viên (Deep Matrix Factorization for Student Performance Prediction)	70
4.2.1	Phương pháp phân rã ma trận	70
4.2.2	Phương pháp phân rã sâu ma trận	71
4.3	Kỹ thuật phân rã sâu ma trận thiên vị trong dự đoán kết quả học tập sinh viên (Deep Biased Matrix Factorization for Student Performance Prediction)	78
4.3.1	Phương pháp phân rã ma trận thiên vị	78
4.3.2	Phương pháp phân rã sâu ma trận thiên vị	78
4.4	Đánh giá kết quả	82
4.4.1	Tập dữ liệu	82
4.4.2	Đánh giá mô hình	83
4.4.3	Tham số thực nghiệm	83
4.4.4	Kết quả thực nghiệm	84
4.5	Tổng kết chương	84
CHƯƠNG 5	TÍCH HỢP CÁC MỐI QUAN HỆ DỮ LIỆU VÀO DỰ ĐOÁN	
	KẾT QUẢ HỌC TẬP CỦA SINH VIÊN	86
5.1	Đặt vấn đề	86
5.2	Phương pháp tích hợp mối quan hệ người dùng	87
5.2.1	Phương pháp phân rã ma trận chưa tích hợp mối quan hệ	87
5.2.2	Phương pháp tích hợp mối quan hệ bạn bè	89
5.2.3	Đánh giá kết quả thực nghiệm phương pháp tích hợp mối quan hệ người dùng	93
5.3	Phương pháp tích hợp mối liên quan các đối tượng	95
5.3.1	Phương pháp tích hợp mối liên quan các môn học	95
5.3.2	Đánh giá kết quả thực nghiệm	100
5.4	Tổng kết chương	102

CHƯƠNG 6	KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN	104
6.1	Kết luận	104
6.2	Hướng phát triển	105
PHỤ LỤC		P1
	XÂY DỰNG ỨNG DỤNG THỬ NGHIỆM	P1

DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT

Viết tắt	Giải thích
ANN	Mạng nơ-ron nhân tạo (Artificial Neural Network)
AUC	Area Under the Curve (Độ đo chính xác)
BMF	Biased Matrix Factorization (Phân rã ma trận thiên vị)
CBF	Content Based Filtering (Lọc nội dung)
CF	Collaborative Filtering (Phương pháp lọc cộng tác)
CRMF	Courses Relationship Matrix Factorization
CRISP-DM	CRoss Industry Standard Process for Data Mining
CTU	Can Tho University (Trường Đại học Cần Thơ)
DL	Deep Learning (Học sâu)
DLMF	Deep Learning Matrix Factorization (Phân rã ma trận sâu)
EDM	Eduational Data Mining (Khai phá dữ liệu giáo dục)
GPA	Grade Point Average (Điểm trung bình)
HF	Hybrid Filtering (Lọc kết hợp)
ITS	Intelligent Tutoring System (Hệ trợ giảng thông minh)
KC	Knowleagde Component (Thành phần kiến thức)
KO	Knowleagde Ontology (Bản thể học kiến thức)
LMS	Learning Managerment System (Hệ thống quản trị học tập)
MAE	Mean Absolute (Độ đo trung bình giá trị tuyệt đối lỗi)
MAP	Mean Average Precision (Độ đo độ chính xác trung bình tuyệt đối)
MF	Matrix Factorization (Phân rã ma trận)
ML	Machine Learning (Học máy)
RMSE	Root Mean Square Error (Lỗi trung bình bình phương gốc)
RS	Recommender System (Hệ thống gợi ý)
SGD	Stochastic Gradient Descent (Giảm dốc ngẫu nhiên)
SVM	Support Vector Machine (Máy học véc-tơ hỗ trợ)

Viết tắt	Giải thích
CSDL	Cơ sở dữ liệu
CTDT	Chương trình đào tạo
CVHT	Cổ vấn học tập
KHHT	Kế hoạch học tập
KHGD	Kế hoạch giảng dạy
RBTV	Ràng buộc toàn vẹn

DANH MỤC CÁC HÌNH VẼ

Hình 1.1	Sự tương đồng giữa hệ thống RS và hệ thống dự đoán kết quả học tập	3
Hình 1.2	Sự tương đồng giữa bài toán dự đoán điểm với bài toán xếp hạng . .	4
Hình 1.3	Các mô hình dự đoán năng lực học tập của sinh viên	5
Hình 1.4	Mô hình nhân tố tiềm ẩn để dự đoán kết quả các môn học cho từng sinh viên cụ thể	5
Hình 1.5	Dữ liệu điểm với ba môn cần dự đoán của sinh viên sv5	6
Hình 1.6	Bảng điểm sau khi dự đoán và hướng gợi ý	7
Hình 2.1	Các hướng tiếp cận truyền thống và xu hướng hiện nay của hệ thống gợi ý	29
Hình 2.2	Tiến trình xử lý của hệ thống gợi ý lọc theo nội dung	31
Hình 2.3	Tiến trình xử lý của hệ tư vấn sử dụng lọc cộng tác	32
Hình 2.4	Các phương pháp lọc kết hợp giữa lọc cộng tác và lọc nội dung . .	34
Hình 2.5	Quy trình khai thác dữ liệu chuẩn CRISP-DM	36
Hình 2.6	Phân rã mô hình dự đoán trên tập dữ liệu ASSISTments	37
Hình 2.7	Mô hình tổng quát cho hệ thống dự đoán	40
Hình 3.1	Ví dụ ma trận đánh giá năng lực sinh viên	44
Hình 3.2	Phương pháp phân chia tập dữ liệu phục vụ cho đánh giá hệ thống gợi ý	46
Hình 3.3	Minh họa kỹ thuật phân rã ma trận	56
Hình 3.4	Minh họa kỹ thuật phân rã ma trận thiên vị	60
Hình 3.5	Kết quả thực nghiệm trên tập dữ liệu CTU	65
Hình 4.1	Ví dụ về mô hình dự đoán năng lực của sinh viên khi làm bài tập . .	69
Hình 4.2	Minh họa kỹ thuật phân rã ma trận	70
Hình 4.3	Mô hình tích hợp kiến trúc học sâu vào phân rã ma trận	72
Hình 4.4	Minh họa kỹ thuật phân rã ma trận thiên vị	78

Hình 4.5	Mô hình tích hợp kiến trúc học sâu vào phân rã ma trận thiên vị . . .	79
Hình 4.6	Mẫu dữ liệu của tập dữ liệu ASSISTments	82
Hình 4.7	Kết quả thực nghiệm trên tập dữ liệu ASSISTment	84
Hình 5.1	Mô hình đồ họa của kỹ thuật phân rã ma trận cơ sở	87
Hình 5.2	Sự chuyển đổi mỗi quan hệ bạn bè thành ma trận mỗi quan hệ . . .	89
Hình 5.3	Mô hình tích hợp mỗi quan hệ người dùng vào kỹ thuật phân rã ma trận	90
Hình 5.4	So sánh độ lỗi RMSE trên 2 tập dữ liệu ở Việt Nam và Quốc tế . . .	95
Hình 5.5	Mô hình chuyển đổi quan hệ môn học thành ma trận	96
Hình 5.6	Mô hình tích hợp mỗi liên quan môn học vào kỹ thuật phân rã ma trận	97
Hình 5.7	Bảng so sánh độ lỗi RMSE của các giải thuật dự đoán	101
Hình 1	Hình kiến trúc hệ thống triển khai	P1
Hình 2	Sơ đồ lớp dữ liệu liên quan đến nhóm sinh viên (users)	P3
Hình 3	Sơ đồ lớp liên quan đến dữ liệu môn học (items)	P4
Hình 4	Sơ đồ lớp liên quan đến dữ liệu điểm (ratings)	P5
Hình 5	Sơ đồ chức năng của hệ thống nền web	P6
Hình 6	Giao diện ứng dụng desktop chức năng dự đoán kết quả học tập sinh viên	P7

DANH MỤC CÁC BẢNG BIỂU

Bảng 2.1	Các giải thuật dự đoán năng lực học tập sinh viên theo hướng dự đoán không cá nhân hóa	20
Bảng 2.2	Một số mã nguồn mở để phát triển hệ thống gợi ý	27
Bảng 3.1	Các tham số thực nghiệm các giải thuật	64
Bảng 4.1	Các tham số của giải thuật tích hợp kiến trúc học sâu thực nghiệm trên tập dữ liệu ASSISTments	83
Bảng 5.1	Tham số thực nghiệm trên tập dữ liệu CTU	94
Bảng 5.2	Tham số thực nghiệm trên tập dữ liệu ASSISTment	94
Bảng 5.3	Các tham số thực nghiệm của các giải thuật trên tập ASSISTments .	101

CHƯƠNG 1 GIỚI THIỆU

Trong chương này, luận án sẽ trình bày sự cần thiết vấn đề gợi ý trong việc cố vấn học tập. Bài toán nghiên cứu được trình bày chi tiết, cũng như tìm ra được những thách thức nghiên cứu. Tiếp theo là đặt ra năm (05) mục tiêu cụ thể, để từ đó đưa ra hướng tiếp cận nghiên cứu của luận án. Đối tượng, phạm vi, phương pháp nghiên cứu cũng được trình bày trong chương này. Luận án đã có bốn (04) đóng góp mới về mặt học thuật cũng như lý luận. Phần cuối của chương này là phần trình bày bố cục của luận án.

1.1 Tính cấp thiết của luận án

Cố vấn học tập (CVHT) là một yếu tố quan trọng trong việc đảm bảo sự thành công của học sinh, sinh viên và ảnh hưởng lớn đến nền giáo dục. Một trong những vấn đề quan trọng của cố vấn học tập là giúp sinh viên xác định mục tiêu học tập và lập các kế hoạch để đạt được mục tiêu đó. Cố vấn học tập cũng có thể giúp sinh viên đánh giá năng lực của mình và đề xuất các hoạt động hoặc khóa học có thể giúp họ phát triển kỹ năng và kiến thức cần thiết. Ngoài ra, cố vấn học tập cũng giúp sinh viên giải quyết các vấn đề cá nhân hoặc học tập, bao gồm cả tình trạng căng thẳng và stress, và cung cấp các tài nguyên để các sinh viên vượt qua những khó khăn đang gặp phải. Các cố vấn học tập cũng có thể giúp sinh viên phát triển các kỹ năng xã hội và nghề nghiệp cần thiết cho sự nghiệp sau này. Chính vì vậy cố vấn học tập (CVHT) có vai trò rất quan trọng trong đào tạo theo tín chỉ, công việc cố vấn học tập không những ảnh hưởng trực tiếp đến sự thành công hay thất bại trong học tập và nghề nghiệp của sinh viên mà còn ảnh hưởng rất lớn đến sự phát triển của xã hội vì giáo dục là quốc sách hàng đầu. Tuy nhiên, công việc của cố vấn học tập gặp nhiều khó khăn, tốn nhiều công sức và thời gian. Vì vậy những nghiên cứu tập trung giải quyết các vấn đề nhằm hỗ trợ CVHT đưa ra những quyết định gợi ý sinh viên một cách chính xác và kịp thời, đang là những hướng nghiên cứu được quan tâm hiện nay.

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] .

Sinh viên muốn lập kế hoạch học tập tốt thì công việc quan trọng nhất là phải lựa chọn môn học phù hợp với năng lực của chính mình. Hiện nay, phần lớn các trường đại học đã triển khai theo học chế tín chỉ nên các sinh viên thường bị lúng túng khi lựa chọn môn học do có nhiều môn được giảng dạy trong một học kỳ. Khi đó, bên cạnh khả năng tự tìm hiểu thì sinh viên sẽ cần đến sự trợ giúp của CVHT. Tuy vậy, để tư vấn cho sinh viên đòi hỏi CVHT phải tra cứu kết quả học tập của từng sinh viên để đưa ra trợ giúp (có phần chủ quan của CVHT) cho mỗi sinh viên, do đó khá tốn thời gian và công sức mà không đạt được hiệu quả cao. Vấn đề là làm sao sử dụng nguồn dữ liệu điểm sinh viên để khai thác, đồng thời sử dụng tiến bộ của khoa học máy tính để phân tích và đưa ra đánh giá hoặc dự đoán một cách chính xác để có thể gợi ý cho sinh viên chọn môn học một cách hiệu quả và tự động thông qua hệ thống.

Để giải quyết các vấn đề cấp thiết này nên nhiều nghiên cứu trong và ngoài nước đã tập trung tìm kiếm và đưa ra các giải pháp ứng dụng khoa học máy tính vào giải quyết các bài toán giáo dục [2] [3] [4] như: dự đoán kết quả học tập sinh viên bằng luật kết hợp, mạng bayes, cây quyết định, lý thuyết tập thô, học sâu mạng nơ-ron đa tầng... Tuy nhiên những nghiên cứu này chỉ có thể dự đoán số đông mà không thể dự đoán cho từng đối tượng cụ thể.

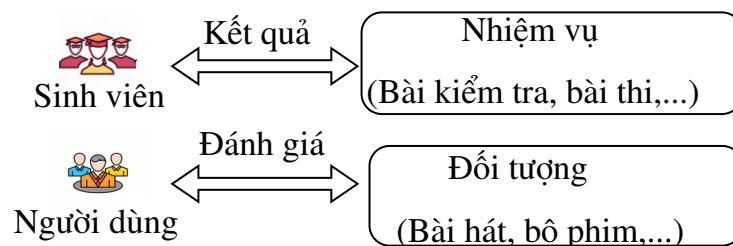
Để giải quyết dự đoán cho cá nhân cụ thể thì phương pháp có hiệu quả và phù hợp nhất là hệ thống gợi ý (Recommender System - RS). Nghiên cứu [5] đã tổng hợp các phương pháp xây dựng hệ thống gợi ý và đề xuất một số phương pháp cải tiến. Với sự phát triển rực rỡ của RS trong nhiều lĩnh vực, nhiều nhà nghiên cứu đã mạnh dạng ứng dụng kỹ thuật này vào khai phá dữ liệu giáo dục nhằm tìm kiếm những giải pháp dự đoán riêng cho từng sinh viên [6] [7]. Các giải thuật gợi ý được áp dụng ban đầu có độ chính xác chưa cao, nhiều thông tin hữu ích về dữ liệu giáo dục chưa được khai thác, các cải tiến theo xu hướng phát triển của kỹ thuật tích hợp cũng chưa được áp dụng. Đây là những bỏ ngỏ cần các nhà khoa học khai thác và nghiên cứu.

Luận án sẽ trình bày tổng quan về bài toán dự đoán năng lực học tập sinh viên, khởi đầu là phân tích và xử lý dữ liệu cho phù hợp với thuật toán, để từ đó đưa ra mô hình dự đoán theo nhiều hướng cải tiến hệ thống gợi ý. Bên cạnh đó luận án cũng trình bày kết quả đánh giá mô hình dự đoán nhằm lựa chọn mô hình tốt nhất và có độ chính xác cao. Cuối cùng là xây dựng hệ thống thử nghiệm nhằm ứng dụng nghiên cứu lý thuyết đi vào thực tiễn.

1.2 Bài toán nghiên cứu

1.2.1 Sự tương quan giữa bài toán xếp hạng và bài toán 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 [8]. Trong ngữ cảnh giáo dục, sinh viên học các môn học sẽ có điểm số trung bình môn học (có thể là thang điểm 10 hoặc thang điểm 4 - điểm chữ). Trong khi đó, ở ngữ cảnh hệ thống gợi ý thì ta có người dùng khi 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 trong hệ thống gợi ý, khi đó việc sinh viên đạt được kết quả (điểm số) khi thực hiện các nhiệm vụ học tập sẽ được giải quyết giống với việc người dùng đánh giá (cho sao, thả tim) sản phẩm khi mua hàng.



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

Hệ thống gợi ý được cấu thành từ ba yếu tố: danh sách người dùng (users - u), danh sách các đối tượng như bài hát, bộ phim, sản phẩm (items - 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 cũng có ba yếu tố: danh sách các sinh viên (students - s), danh sách các môn học (courses - c) và điểm số (gradings - g). Như vậy, việc dự đoán đánh giá

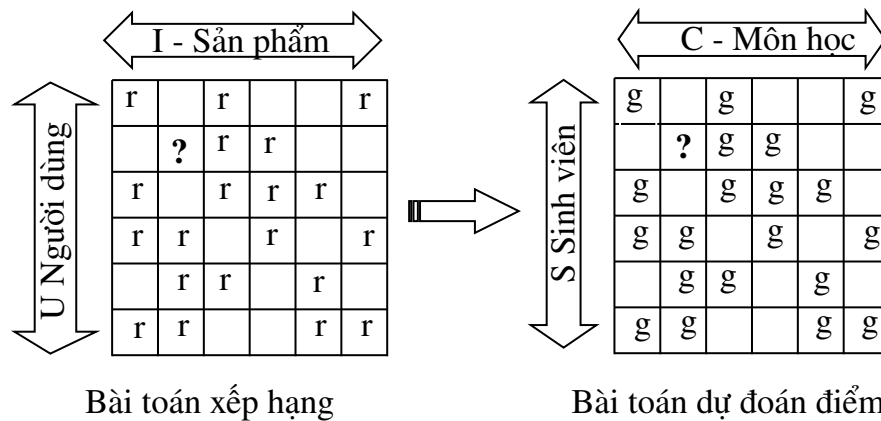
của người dùng trong bài toán xếp hạng (Rating Prediction) của RS sẽ tương đương với bài toán dự đoán năng lực học tập (điểm số) sinh viên (Student Performance Prediction). Hình 1.2 thể hiện sự ánh xạ từ bài toán xếp hạng sang bài toán dự đoán điểm. Tương tự, sự tương đồng các khái niệm được biểu diễn như sau:

Người dùng (u) $\rightarrow$ Sinh viên (s)

Sản phẩm (i) $\rightarrow$ Môn học (c)

Đánh giá (r) $\rightarrow$ Điểm (g)

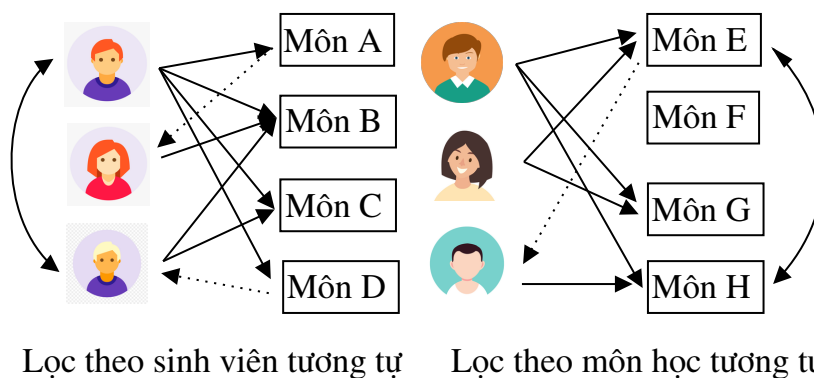
Từ sự tương đồng này, ta có thể giải bài toán dự đoán điểm học tập của sinh viên theo các cách của bài toán xếp hạng trong hệ thống gợi ý.



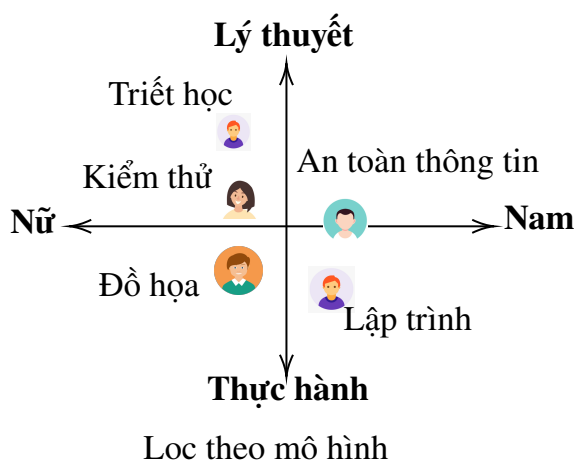
Hình 1.2: Sự tương đồng giữa bài toán dự đoán điểm với bài toán xếp hạng

1.2.2 Bài toán dự đoán kết quả học tập của sinh viên

Do bài toán xếp hạng của hệ thống gợi ý là dạng dự đoán cho từng cá nhân cụ thể để đưa ra tư vấn, nên mô hình dự đoán điểm học tập sinh viên theo hướng tiếp cận cá nhân hóa cũng dựa trên ý tưởng của các phương pháp lọc cộng tác (Colaborative Fittering - CF) trong hệ thống gợi ý. Hình 1.3 minh họa bài toán dự đoán năng lực sinh viên theo ba nhóm: Lọc theo sinh viên tương tự (Student-based Filtering), lọc theo môn học tương tự (Course-based Filtering) và lọc theo mô hình (Model-based Filtering). Lọc theo sinh viên tương tự sẽ dựa vào năng lực tương đồng của hai sinh viên và lấy kết quả của một sinh viên đã học để dự đoán kết quả cho sinh viên chưa học. Tương tự, lọc theo môn học sẽ dựa vào môn học giống nhau để dự đoán gợi ý cho sinh viên khi đã học một môn tương tự. Lọc theo mô hình sẽ dựa vào mô hình nhân tố tiềm ẩn để dự đoán kết quả các môn học cho từng sinh viên cụ thể (Hình 1.4).



Hình 1.3: Các mô hình dự đoán năng lực học tập của sinh viên



Hình 1.4: Mô hình nhân tố tiềm ẩn để dự đoán kết quả các môn học cho từng sinh viên cụ thể

1.2.3 Gợi ý lựa chọn môn học tự chọn

Đầu tiên chúng ta cần sử dụng và khai thác thông tin điểm số của sinh viên bằng các kỹ thuật khai thác dữ liệu từ đó đưa ra được kết quả dự đoán cho từng sinh viên cụ thể theo từng học kỳ, trước khi sinh viên tham gia lập kế hoạch học tập ở đầu mỗi học kỳ. Từ những dự đoán kết quả của sinh viên với các môn chưa học, làm cơ sở cho hệ thống cung cấp những gợi ý nên chọn môn học nào trong những môn tự chọn là phù hợp nhưng vẫn đảm bảo áp dụng đúng với ràng buộc của chương trình đào tạo như: những môn học trước, môn học tiên quyết, số lượng các môn học cần phải chọn trong nhóm, v.v...

Ví dụ: Có năm 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 (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. Với điều kiện trong chương trình đào tạo là sinh viên phải chọn 2 môn trong 3 môn học tự chọn (n1,n2,n3). Khi đó, cố vấn học tập là người hướng dẫn và tư vấn cho sinh viên nên chọn những môn học nào là phù hợp, họ phải tìm bảng điểm của mỗi sinh viên cũng như những thông tin về môn học rồi đưa ra như tư vấn có phần chủ quan, nếu như có hệ thống gợi ý thì công việc này của cố vấn học tập thật sự được nhẹ nhàng hơn, dành nhiều thời gian cho những tư vấn khác. Như vậy, lúc này hệ thống cần gợi ý cho sinh viên sv5 là nên học 2 môn nào trong 3 môn: Môn n1, Môn n2, Môn n3.

		C - Course (Môn học)						
		Môn	Môn		Môn	Môn	Môn	
		1	2		n	n1	n2	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

Sau khi chạy giải thuật khai phá dữ liệu và thực hiện dự đoán cho tất cả sinh viên sẽ 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.

Trở lại ví dụ trên, hệ thống cần gợi ý 2 môn học tự chọn trong nhóm 3 môn tự chọn (sao cho tích lũy đủ số tín chỉ) 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)								
S - Student (Sinh viên)		Môn 1	Môn 2		Môn n	Môn n1	Môn n2	Môn n3
	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

Thách thức thứ nhất là 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 để tư vấn cho từng sinh viên cụ thể một cách phù hợp nhất, trong khi phần lớn các nghiên cứu hiện nay chỉ tập trung vào dự đoán cho nhóm các người dùng, hoặc đưa ra các quy luật chung.

Thách thức thứ hai là 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.

Thách thức thứ ba là thiếu xem xét các mối quan hệ người dùng và mối liên quan giữa các môn học: Các sinh viên có cùng môi trường sẽ ảnh hưởng nhau, cũng như những môn học có nội dung liên quan nhau, đây là thuận lợi cho việc khai thác nhằm nâng cao hiệu quả dự đoán.

Thách thức thứ tư là độ chính xác của dự đoán còn có thể nâng cao bằng các mô hình tích hợp: Hiện nay, khoa học càng phát triển, những phát minh đã thành công ở lĩnh vực này, thì có thể áp dụng thành công ở lĩnh vực khác. Thật vậy, có nhiều nghiên cứu về khai phá dữ liệu giáo dục tuy nhiên độ chính xác vẫn cần được cải thiện, nên cần nghiên cứu những thuật toán mạnh hơn, hiệu quả hơn. Vì thế, việc tìm các giải pháp nâng cao hiệu quả dự đoán bằng việc tận dụng hay tích hợp được những kiến trúc khác để giải quyết vấn đề hiệu quả hơn, là một trong những thách thức nghiên cứu.

1.4 Các vấn đề nghiên cứu

Từ việc phân tích các thách thức, vấn đề còn bỏ ngỏ của các nghiên cứu liên quan, luận án xác định một số vấn đề cần nghiên cứu bao gồm:

- Vấn đề dự đoán năng lực học tập của sinh viên theo hướng tiếp cận 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 hiệu quả dự đoán.
- Vấn đề xét mối liên quan về nội dung kiến thức giữa các môn học nhằm đưa ra dự đoán chính xác hơn.
- 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 từng sinh viên cụ thể.

Mục tiêu 1: Đề xuất quy trình giải pháp gợi ý lựa chọn môn học của cổ vấn học tập theo hướng cá nhân hóa: Ứng dụng các giải thuật cơ bản trong hệ thống gợi ý vào dự đoán năng lực học tập của sinh viên như: lọc cộng tác theo môn học tương tự (Course-based Filtering), lọc cộng tác theo sinh viên tương tự (Student-based Filtering), lọc cộng

tác theo mô hình phân rã ma trận (Matrix Factorization - MF).

Mục tiêu 2: Đề xuất giải pháp giải quyết vấn đề không đồng nhất trong đánh giá năng lực học tập của sinh viên: Giải quyết vấn đề dự đoán khi dữ liệu đánh giá không đồng nhất (đã trình bày trong thách thức nghiên cứu thứ hai) bởi giải thuật phân rã ma trận thiên vị (Biased Matrix Factorization - BMF) và so sánh với các giải thuật cơ sở.

Mục tiêu 3: Đề xuất phương pháp tích hợp mối quan hệ người dùng vào mô hình phân rã ma trận nhằm nâng cao độ chính xác trong dự đoán kết quả học tập của sinh viên: Đề xuất xử lý dữ liệu mối quan hệ bạn bè và mô hình tích hợp dữ liệu mối quan hệ bạn bè này vào mô hình dự đoán kết quả học tập sinh viên. Xây dựng thuật toán (Social Matrix Factorization - SocialMF) được thực nghiệm trên tập dữ liệu giáo dục và đánh giá phương pháp tích hợp mối quan hệ người dùng so với các phương pháp chưa tích hợp.

Mục tiêu 4: Đề xuất phương pháp tích hợp mối liên quan nội dung của các môn học vào mô hình dự đoán nhằm nâng cao độ chính xác trong dự đoán kết quả học tập của sinh viên: Đề xuất xử lý dữ liệu mối liên quan của các môn học vào mô hình tích hợp dữ liệu mối liên quan môn học này vào mô hình dự đoán kết quả học tập sinh viên. Xây dựng thuật toán (Courses Relationship Matrix Factorization - CRMF), thực nghiệm và đánh giá phương pháp tích hợp mối liên quan của các môn học so với các phương pháp chưa tích hợp.

Mục tiêu 5: Đề xuất ứng dụng kiến trúc học sâu vào mô hình dự đoán nhằm nâng cao hiệu quả dự đoán kết quả học tập sinh viên: Đề xuất phương pháp nâng cao độ chính xác dự đoán kết quả học tập của sinh viên theo hướng tiếp cận tích hợp kiến trúc học sâu vào kỹ thuật phân rã ma trận (Deep Matrix Factorization - DMF). Tiếp nối theo là phương pháp phân rã sâu ma trận thiên vị (Deep Biased Matrix Factorization - DBMF) cũng được đề xuất nhằm cải tiến mô hình dự đoán.

1.6 Đối tượng, phương pháp và phạm vi nghiên cứu

Đối tượng nghiên cứu chính là các mô hình dự đoán năng lực học tập của sinh viên, các giải pháp gợi ý cá nhân hóa, và các tập dữ liệu hỗ trợ.

Phạm vi nghiên cứu của luận án là tập trung vào các phương pháp xây dựng và cải

tiền mô hình dự đoán điểm học tập của sinh viên để từ đó đưa ra gợi ý lựa chọn môn học phù hợp cho từng cá nhân cụ thể, thực nghiệm trên hai tập dữ liệu ở Việt Nam (tập dữ liệu điểm của trường Đại học Cần Thơ) và Quốc tế (tập dữ liệu ASSITSment).

Để thực hiện nghiên cứu, chúng tôi dùng phương pháp nghiên cứu lý thuyết để phân tích và tổng hợp các nghiên cứu có liên quan đến nội dung nghiên cứu từ các tài liệu tham khảo: sách, bài báo được công bố trên các tạp chí, kỷ yếu, hội thảo hội nghị có uy tín để đề xuất các phương pháp mới. Phương pháp nghiên cứu cơ bản trong luận án này là đề xuất tích hợp nhiều dữ liệu và nhiều phương pháp cải tiến mới nhất.

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. Đóng góp cụ thể được tổng hợp từ các công bố [CT1] bao gồm: Phát huy điểm mạnh của hệ thống gợi ý là đưa ra những dự đoán và gợi ý cá nhân hóa. Đồng thời, phát hiện sự tương đồng giữa hai bài toán xếp hạng trong hệ thống gợi ý với bài toán dự đoán năng lực học tập của sinh viên. Luận án đã đề xuất quy trình gợi ý môn học tự chọn trong việc cố vấn học tập theo hướng tiếp cận cá nhân hóa. Bên cạnh đó, vấn đề dữ liệu đánh giá không đồng nhất (thách thức nghiên cứu thứ hai) cũng được giải quyết bởi giải thuật phân rã ma trận thiên vị.

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 nhằm nâng cao hiệu quả dự đoán. Đóng góp cụ thể được tổng hợp từ các công bố [CT4][CT5] bao gồm: Độ chính xác của các giải thuật trong hệ thống gợi ý luôn được cải tiến và nâng cao hiệu quả. Luận án đã đề xuất tích hợp kiến trúc học sâu vào kỹ thuật phân rã ma trận nhằm huấn luyện ma trận qua nhiều tầng cho đến khi đạt được mô hình tối ưu nhất. Thêm vào đó, luận án cũng thực nghiệm mô hình với đề xuất áp dụng kiến trúc học sâu cho phân rã ma trận thiên vị. Kết quả thí nghiệm cho thấy cả hai việc tích hợp kiến trúc học sâu vào phân rã ma trận hay phân rã ma trận thiên vị đã cải thiện được độ chính xác của dự đoán.

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. Đóng góp cụ thể được tổng hợp từ công bố [CT2]: Với nguồn thông tin dồi dào 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. 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 để bổ sung thông tin, làm cho việc dự đoán đạt kết quả tốt hơn. Luận án đã đề xuất việc tích hợp mối quan hệ bạn bè cùng lớp để tích hợp vào kỹ thuật phân rã ma trận nhằm nâng cao độ chính xác của dự đoán. Kết quả thực nghiệm cho thấy mô hình đề xuất này đã được cải thiện và tốt hơn.

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. Đóng góp cụ thể được tổng hợp từ công bố [CT3]: Tiếp nối sự thành công của việc tích hợp mối quan hệ người dùng trong hệ thống gợi ý (việc tích hợp phía người dùng - user). Luận án đã đề xuất tích hợp các mối liên quan giữa kiến thức cũng như nội dung của môn học (tích hợp phía mục tin - item) vào kỹ thuật phân rã ma trận. Một lần nữa, kết quả thực nghiệm cho thấy việc tích hợp này hoàn toàn khả thi và mang lại hiệu quả cho mô hình dự đoán này.

1.8 Bố cục của luận án

Chương 1 giới thiệu tính cần thiết cũng như ý nghĩa của hướng nghiên cứu. Bắt đầu từ bài toán nghiên cứu được trình bày chi tiết, cũng như tìm ra được những thách thức nghiên cứu. Tiếp theo chúng tôi xác định mục tiêu và hướng tiếp cận của luận án để giải quyết các thách thức của bài toán nghiên cứu. Cuối cùng chúng tôi trình bày các đóng góp của luận án và giới thiệu tóm tắt cấu trúc của luận án.

Chương 2 giới thiệu các nghiên cứu liên quan và thảo luận các hướng tiếp cận để nâng cao hiệu quả của dự đoán. Bên cạnh đó, luận án còn trình bày về hệ trợ giảng thông minh, hệ thống gợi ý và hệ thống khai thác dữ liệu giáo dục.

Chương 3 trình bày chi tiết bài toán dự đoán kết quả học tập của sinh viên theo hai hướng tiếp cận: dự đoán không cá nhân hóa và dự đoán cá nhân hóa, từ đó đề xuất mô hình gợi ý lựa chọn môn học nhằm hỗ trợ cố vấn học tập. đồng thời đề xuất áp dụng kỹ thuật phân rã ma trận thiên vị (BMF) nhằm giải quyết bài toán đánh giá không đồng nhất

(thách thức nghiên cứu thứ hai) [CT1].

Chương 4 đề xuất phương pháp tích hợp kiến trúc học sâu vào kỹ thuật phân rã ma trận (DMF), cũng như phân rã sâu ma trận thiên vị (DBMF) vào bài toán dự đoán kết quả học tập của sinh viên. Nội dung trình bày chi tiết các giải thuật phân rã ma trận cơ sở (MF và BMF) trước khi tích hợp và tiếp theo là cách thức tích hợp mô hình học sâu cho hai kỹ thuật cơ sở này. Phần sau cùng là kết quả thí nghiệm cũng như những hướng nghiên cứu tiếp theo [CT4][CT5].

Chương 5 đề xuất các phương pháp tích hợp dữ liệu bổ sung nhằm nâng cao hiệu quả dự đoán. Trình bày các phương pháp dự đoán kết quả học tập của sinh viên sử dụng kỹ thuật hệ thống gợi ý. Nội dung được trình bày từ các phương pháp cơ sở của hệ thống gợi ý đến kỹ thuật tích hợp mối quan hệ người dùng vào mô hình phân rã ma trận (SocialMF) nhằm nâng cao độ chính xác của dự đoán kết quả học tập của sinh viên, kết quả thí nghiệm cho thấy kết quả đạt độ chính xác cao hơn khi ứng dụng việc tích hợp này. Nối tiếp sự thành công, chúng tôi đã mở rộng việc tích hợp với giả thuyết tương tự cho những môn học có nội dung tương đồng nhau (CRMF). Hai phương pháp đề xuất được tổng hợp từ hai công trình đã công bố [CT2][CT3].

Chương 6 trình bày kết luận tóm tắt các đóng góp của luận án cũng như những hạn chế mà luận án chưa giải quyết và đưa ra một số hướng nghiên cứu tiếp theo.

Phần phụ lục trình bày kiến trúc tổng thể trong việc xây dựng hệ thống gợi ý lựa chọn môn học của sinh viên bao gồm: thiết kế cơ sở dữ liệu lưu trữ hệ thống, mô hình tích hợp giải thuật gợi ý vào hệ thống, và thiết kế các giao diện nền web và ứng dụng.

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

Mục tiêu chính của chương này là trình bày cơ sở lý thuyết và các nghiên cứu liên quan đến bài toán dự đoán kết quả học tập của sinh viên từ đó đưa ra gợi ý lựa chọn môn học phù hợp.

Nội dung chính của chương này là trình bày chi tiết những nghiên cứu liên quan về hệ thống thông tin phù hợp cũng như các phương pháp khai thác dữ liệu tiên tiến, được phân thành hai nhóm: (1) Các nghiên cứu về hệ thống thông tin bao gồm: hệ thống gợi ý, hệ trợ giảng thông minh, hệ thống dự đoán năng lực học tập sinh viên; (2) Các nghiên cứu về việc cải tiến phương pháp khai thác dữ liệu như: các phương pháp nâng cao hiệu quả hệ thống gợi ý, các phương pháp tích hợp mối quan hệ dữ liệu, các phương pháp cải tiến mô hình dự đoán. Bên cạnh đó, chương này cũng đưa ra những nhận xét đánh giá và xác định các thách thức cần nghiên cứu của bài toán. Từ đó, luận án xác định các phương pháp khai thác dữ liệu giáo dục phù hợp cũng như đề xuất quy trình để xây dựng hệ thống dự đoán năng lực học tập của sinh viên. Các hướng nghiên cứu cụ thể được trình bày ở các chương sau của luận án.

2.1 Các nghiên cứu liên quan

2.1.1 Các nghiên cứu về khai thác dữ liệu giáo dục

Khai thác dữ liệu giáo dục (Education Data Mining - EDM) là một lĩnh vực nghiên cứu nhằm áp dụng các kỹ thuật khai thác dữ liệu trên dữ liệu của hệ thống giáo dục, với mục đích khám phá và tìm ra các thông tin tiềm ẩn để giải quyết các câu hỏi và vấn đề chưa được giải đáp trong lĩnh vực giáo dục. Từ đó, các cải tiến có thể được đưa ra để nâng cao chất lượng giảng dạy và học tập.

EDM được coi là một kho tàng vô giá trong lĩnh vực khai thác dữ liệu bởi vì nó chứa đựng rất nhiều dữ liệu đã tích lũy được trong nhiều năm và đa dạng về chủ đề.

Các phương pháp thường được sử dụng trong các nghiên cứu EDM là các phương pháp truyền thống [9]. Ví dụ, cây quyết định được sử dụng để khai thác dữ liệu nhằm hỗ trợ sinh viên đăng ký các khóa học hoặc dự báo khả năng học tập của họ dựa trên các thuộc tính tuyển sinh [10]. Nghiên cứu [11] đã so sánh các phương pháp được sử dụng trong EDM để đánh giá và đề xuất các phương pháp phù hợp với các bài toán cụ thể. Nhìn chung, nghiên cứu về EDM có thể được chia thành hai hướng tiếp cận:

- EDM được sử dụng để phân tích hành vi của người học.
- EDM được sử dụng để dự báo và đánh giá khả năng của người học.

2.1.1.1 Các nghiên cứu phân tích hành vi người học

Việc đánh giá hành vi và thái độ của người học là một yếu tố quan trọng để tối ưu hóa phương pháp giảng dạy. Trong các nghiên cứu về phân tích hành vi của người học [12], thường sử dụng các kỹ thuật phân loại tuần tự và gom nhóm để phân tích việc sử dụng tài nguyên học tập của sinh viên, quá trình học tập, thực hành, làm bài tập và kiểm tra. Việc phân tích hành vi trong thời gian thực giúp cung cấp thông tin phản hồi kịp thời cho giảng viên và người học, từ đó giúp nâng cao quá trình giảng dạy và học tập. Các nghiên cứu về hành vi của người học đã đóng góp đáng kể vào việc cải thiện môi trường học tập. Tuy nhiên, những nghiên cứu này vẫn phân loại, gom nhóm theo các kỹ thuật truyền thống, chưa quan tâm vấn đề dự đoán và phân tích cá nhân hóa.

2.1.1.2 Các nghiên cứu dự báo và đánh giá khả năng của người học

Hiện nay, việc khai thác dữ liệu trong giáo dục đóng vai trò quan trọng trong việc phát triển hệ thống giáo dục. Các ứng dụng khai thác dữ liệu trong giáo dục như EDM, đã được sử dụng chủ yếu trong các hệ thống quản trị học tập (Learning Management Systems - LMS) và hệ thống trợ giảng thông minh (Intelligent Tutoring System- ITS) như được đề cập trong tài liệu [9]. Một trong những hệ thống LMS tiêu biểu và phổ biến nhất hiện nay là Moodle, nghiên cứu đã sử dụng kỹ thuật khai thác dữ liệu trên hệ thống Moodle để tổng hợp các mô hình phân loại ứng dụng và thực hiện dự đoán năng lực của sinh viên [13] [14]. Hệ thống này được thiết kế để trích xuất dữ liệu cần thiết một cách tự động và hỗ trợ cho hoạt động giảng dạy của giảng viên. Một nghiên cứu khác đã sử dụng cây quyết định để khai thác dữ liệu từ thông tin truy cập hệ thống Moodle của sinh

viên [15]. Nhờ đó, nghiên cứu này đã xác định và xếp hạng chính xác kết quả thi cuối khóa của sinh viên thông qua việc tham gia các lớp học trên Moodle. Ngoài việc dự đoán kết quả học của sinh viên, khai thác dữ liệu trên hệ thống Moodle còn giúp phát hiện những truy cập không thường xuyên của sinh viên. Nhiều thí nghiệm khai thác dữ liệu các hệ thống e-learning cũng đã được thực hiện để dự đoán điểm cuối khóa của sinh viên và phân loại sinh viên theo nhiều ứng dụng khác nhau, ví dụ như phát hiện các nhóm sinh viên có cùng đặc trưng, xác định nhóm người học thụ động để đề xuất hướng khắc phục, dự đoán và phân loại nhóm sinh viên có thường xuyên sử dụng hệ thống tài liệu thông minh [16].

Các nghiên cứu của EDM đã áp dụng nhiều kỹ thuật khai thác dữ liệu như phân tích nhân tố, hồi quy logistic, cây quyết định, máy hỗ trợ véc-tor (SVM), và mạng Bayes để xây dựng các mô hình khai thác dữ liệu có thể giúp dự đoán kết quả học tập của sinh viên. Ngoài việc dự đoán kết quả, các nghiên cứu cũng phân tích các kết quả dự đoán để tìm ra những cảnh báo sớm, từ đó đưa ra các gợi ý hành động khắc phục cho các cơ sở giáo dục [17].

Bài toán này cũng là một thành phần quan trọng trong hệ trợ giảng thông minh (ITS), Dominguez và đồng nghiệp đã phát triển một hệ thống tiếp nhận phản hồi từ sinh viên và theo dõi chia sẻ tài liệu học của họ. Kết quả cho thấy, những sinh viên tham gia vào hệ thống và lưu lại trong thời gian dài đã đạt được kết quả tốt hơn so với những sinh viên không tham gia [18]. Nghiên cứu mới đây đã phân tích tương tác của sinh viên với các bài giảng được ghi bằng kỹ thuật khai thác dữ liệu giáo dục và tìm thấy sự khác biệt giữa các báo cáo bằng lời nói của người học và thực tế sử dụng, cho thấy những sinh viên này có kết quả thi tốt hơn [19]. Với tầm quan trọng của việc dự đoán và đánh giá năng lực người học, nhiệm vụ này được coi là rất cần thiết, tuy nhiên các nghiên cứu này vẫn sử dụng các kỹ thuật truyền thống và hạn chế về mặt dữ liệu thực nghiệm.

2.1.2 Các nghiên cứu về hệ thống gợi ý

Trong suốt hơn hai thập kỷ qua, hệ thống gợi ý đã được nghiên cứu và áp dụng rộng rãi. Các nghiên cứu trong lĩnh vực này rất đa dạng và có thể phân loại thành ba hướng chính: nghiên cứu về dữ liệu sử dụng trong hệ thống gợi ý, đề xuất và cải tiến các phương pháp gợi ý, và đánh giá hiệu quả của hệ thống gợi ý. Tuy đã đạt được nhiều thành công,

nhưng tất cả các hướng nghiên cứu này vẫn đang tiếp tục được phát triển để đáp ứng sự đa dạng trong các lĩnh vực ứng dụng, nhu cầu của người dùng và sự tiến bộ của công nghệ. Trong số đó, hướng nghiên cứu cải tiến phương pháp gợi ý vẫn đóng vai trò quan trọng và được quan tâm hàng đầu.

2.1.2.1 Các nghiên cứu về dữ liệu theo định dạng hệ thống gợi ý

Các nghiên cứu về dữ liệu sử dụng trong hệ thống gợi ý có thể được chia thành ba nhóm chính: thu thập dữ liệu, biểu diễn dữ liệu, và xây dựng cũng như công khai các tập dữ liệu hỗ trợ cho các hoạt động thực nghiệm.

Thu thập dữ liệu có thể được thực hiện tường minh hoặc không tường minh. Hiện nay, các nghiên cứu về tiền xử lý dữ liệu ngày càng được quan tâm để cải thiện độ chính xác của dự đoán. Nghiên cứu [20] đã tổng hợp các phương pháp từ việc thu thập dữ liệu, biểu diễn, xử lý nhiễu, rút gọn và bảo mật dữ liệu. Các nghiên cứu khác đã đề xuất các phương pháp thu thập dữ liệu như được đề cập trong [21] và [22]. Để tăng hiệu quả trong hoạt động gợi ý, dữ liệu cần được biểu diễn một cách phù hợp. Các dạng biểu diễn dữ liệu phổ biến có thể là không gian véc-tơ hoặc đồ thị [23]. Nhiều kỹ thuật thu giảm chiều số được sử dụng như phân rã giá trị riêng (singular value decomposition) và phân tích thành phần chính (principle component analysis). Để hỗ trợ các nhà nghiên cứu kiểm tra và so sánh các phương pháp gợi ý, nhiều nghiên cứu đã tập trung vào việc xây dựng và công khai các tập dữ liệu trên các lĩnh vực khác nhau.

Tổng quan về nghiên cứu dữ liệu không chỉ có ý nghĩa đối với hệ thống gợi ý mà còn với nhiều lĩnh vực khác như khai thác dữ liệu hay học máy. Vì vậy, các nhà thiết kế và phát triển hệ thống gợi ý có thể tận dụng những kết quả nghiên cứu dữ liệu trong các lĩnh vực khác và điều chỉnh để áp dụng vào hệ thống của mình. Tuy nhiên, dữ liệu về giáo dục hiện tại vẫn chưa đa dạng đủ để giúp cải tiến các thuật toán khai thác hiệu quả. Việc tìm kiếm, thu thập và xử lý dữ liệu phù hợp có thể đóng góp vào việc nâng cao chất lượng và độ chính xác của hệ thống gợi ý, giúp cải thiện trải nghiệm cho người dùng.

2.1.2.2 Các nghiên cứu đề xuất và cải tiến các phương pháp gợi ý

Nghiên cứu trong lĩnh vực gợi ý đang tập trung vào ba nhóm chính: Nhóm một tập trung vào việc đề xuất mô hình và giải thuật gợi ý mới; Nhóm hai tập trung vào việc cải

tiền các mô hình và giải thuật gợi ý; và Nhóm ba tập trung vào việc khảo sát và đánh giá các phương pháp gợi ý để xác định những chủ đề mới cũng như những thách thức quan trọng cần được nghiên cứu. Các kỹ thuật cơ bản đã được đề xuất để xây dựng các hệ thống gợi ý bao gồm lọc cộng tác, dựa trên tri thức và dựa trên nội dung [24]. Ngoài ra, cũng đã có nhiều phương pháp gợi ý mới được đề xuất để phù hợp với sự phát triển của công nghệ web, internet, thiết bị cảm biến, thiết bị di động và mạng xã hội. Mỗi phương pháp gợi ý lại được xây dựng từ nhiều mô hình và giải thuật khác nhau.

Có rất nhiều kỹ thuật khai thác dữ liệu và học máy được áp dụng vào lĩnh vực gợi ý [21], bao gồm sự phân lớp bằng láng giềng gần nhất, cây quyết định, phân lớp mô hình Bayes, máy véc-tơ hỗ trợ, mạng nơ ron; sự phân cụm bằng k-mean; khai thác luật kết hợp; mô hình hồi quy; các phương pháp học máy có giám sát và không được giám sát và kỹ thuật học sâu (deep learning). Tuy nhiên, nhiều nghiên cứu gần đây đã tập trung vào việc cải tiến các giải thuật và mô hình để nâng cao hiệu suất và đạt được kết quả gợi ý tốt hơn. Các nghiên cứu mới nhất về hệ thống gợi ý, chẳng hạn như [25] [26] [27] [28] [22], đã chỉ ra một số chủ đề mới hoặc vấn đề then chốt như đảm bảo tính riêng tư, tích hợp thông tin từ mạng xã hội, mối quan hệ giữa các đối tượng, và các kỹ thuật tiên tiến hiện nay. Tuy nhiên, các nghiên cứu này chưa được ứng dụng vào lĩnh vực giáo dục.

2.1.2.3 Các nghiên cứu về đánh giá hệ thống gợi ý

Các nghiên cứu về đánh giá hệ thống gợi ý cung cấp cho người làm nghiên cứu và người làm thực tiễn những công cụ và phương pháp để đánh giá và cải thiện chất lượng của các hệ thống gợi ý. Các phương pháp này được phân thành ba nhóm chính:

Nhóm thứ nhất tập trung vào khảo sát và tổng kết các phương pháp, kỹ thuật và độ đo được sử dụng để đánh giá hệ thống gợi ý. Các nghiên cứu như [24] đã cung cấp cho người làm nghiên cứu và người làm thực tiễn một cái nhìn tổng quan về các phương pháp và những vấn đề chưa được giải quyết. Ngoài ra, một số nghiên cứu còn cung cấp các chỉ dẫn thực tế để đánh giá các hệ thống gợi ý, chẳng hạn như chỉ dẫn đánh giá các hệ thống gợi ý được cá nhân hóa.

Nhóm thứ hai của các nghiên cứu về đánh giá hệ thống gợi ý tập trung vào phát triển công cụ đánh giá hệ thống gợi ý, ví dụ như thư viện recommenderlab [29] hoặc các thư viện khác chứa các độ đo đánh giá được viết bằng nhiều ngôn ngữ lập trình khác nhau.

Nhóm thứ ba trong các nghiên cứu về đánh giá hệ thống gợi ý tập trung vào đề xuất các phương pháp so sánh các hệ thống gợi ý theo nhiều khía cạnh khác nhau, bao gồm giải thuật hay các đặc trưng chất lượng. Ngoài ra, những đề xuất sử dụng các độ đo khác để đánh giá hệ thống gợi ý (ngoài các độ đo về tính chính xác) cũng được thảo luận trong nhóm này.

Các nghiên cứu về đánh giá hệ thống gợi ý rất hữu ích cho các thiết kế hệ thống ở nhiều lĩnh vực, tuy nhiên lĩnh vực giáo dục vẫn chưa được quan tâm.

2.1.2.4 Các hướng tiếp cận để nâng cao hiệu quả hệ thống gợi ý

Hướng tiếp cận để nâng cao hiệu quả hệ thống gợi ý có nhiều phương pháp thông thường như thu thập và phân tích dữ liệu, sử dụng thuật toán máy học, đánh giá và kiểm tra chất lượng hệ thống: Thu thập dữ liệu đầy đủ và chính xác về hoạt động của người dùng. Phân tích dữ liệu này để hiểu được những gì người dùng thích và không thích, thói quen mua sắm, ưa thích sản phẩm, v.v. Sử dụng các công cụ phân tích dữ liệu để tìm kiếm mẫu. Ngoài ra, còn sử dụng các thuật toán máy học như bộ lọc cộng tác, phân tích lân cận, và học sâu đều có thể được sử dụng để tạo ra các gợi ý tốt hơn. Thuật toán này có thể học hỏi các mô hình từ dữ liệu, từ đó tạo ra các gợi ý dựa trên những gì người dùng đã làm trước đó. Các phương pháp đánh giá và kiểm tra chất lượng hệ thống gợi ý như đánh giá độ chính xác, độ đa dạng và độ phù hợp đều cần được sử dụng để đảm bảo hệ thống đang hoạt động tốt, các kỹ thuật kiểm tra phân tích để so sánh hiệu quả giữa các phương pháp gợi ý khác nhau.

Các phương pháp truyền thống để xây dựng hệ thống gợi ý bao gồm lọc theo nội dung, lọc cộng tác và lọc kết hợp, đều tập trung vào ba loại thông tin đầu vào bao gồm thông tin người dùng, thông tin mục tin và đánh giá phản hồi của người dùng về các mục tin. Mục đích của việc khai thác thông tin này là để đưa ra các gợi ý mục tin mới phù hợp với người dùng hiện tại. Tuy nhiên, chất lượng của hệ thống gợi ý phụ thuộc vào hiệu quả của phương pháp lọc thông tin cơ bản trên ba loại thông tin này. Mặc dù các phương pháp truyền thống có ưu điểm và hạn chế nhất định, chúng tôi cần tiếp tục nghiên cứu để giải quyết các vấn đề này. Hiện nay, các nghiên cứu về hệ thống gợi ý đang tập trung vào hai xu hướng chính: Một, cải tiến các phương pháp lọc tin truyền thống trong hệ thống gợi ý; Hai, mở rộng các phương pháp gợi ý truyền thống để tích hợp thêm các nguồn

thông tin khác [30].

Các nhà nghiên cứu hiện đang tập trung vào xu hướng đầu tiên, đó là cải tiến các phương pháp lọc tin truyền thống trong hệ thống gợi ý. Trong đó, một điểm đáng chú ý là sự quan tâm đến việc tích hợp mô hình học sâu vào phương pháp phân rã ma trận (Deep Matrix Factorization) để cải thiện chất lượng của hệ thống gợi ý.

Cụ thể, một số nghiên cứu về hệ thống gợi ý theo các phương pháp cải tiến được đề cập đến như: Hợp nhất lọc cộng tác và lọc nội dung để giải quyết bài toán gợi ý [58], áp dụng mô hình học sâu cho lọc kết hợp để giải quyết bài toán gợi ý [31], kết hợp các phương pháp lọc cộng tác khác nhau dựa vào giải thuật phân loại đa lớp để đưa ra gợi ý hay khai thác những tác nhân tiềm ẩn của người dùng và mục tin cho hệ thống gợi ý cộng tác dựa trên mô hình mạng nơ-ron nhân tạo [32]... Các phương pháp lọc kết hợp mới này được đánh giá là cho độ chính xác tốt, tuy nhiên hầu hết đang tập trung vào một số bài toán gợi ý cụ thể hoặc có độ phức tạp tính toán tương đối lớn.

Xu hướng thứ hai trong nghiên cứu hệ thống gợi ý là sử dụng các nguồn thông tin phong phú và đa dạng hơn để cải thiện quá trình gợi ý. Bên cạnh việc khai thác các thông tin cơ bản như người dùng, mục tin và phản hồi của người dùng về mục tin, cộng đồng nghiên cứu đang khám phá việc tích hợp thêm các nguồn thông tin hữu ích khác thu được từ hệ thống, như dữ liệu về hành vi người dùng, thông tin về sản phẩm liên quan và nhiều hơn nữa. Việc sử dụng các nguồn thông tin đa dạng này có thể giúp cải thiện độ chính xác và độ phù hợp của hệ thống gợi ý.

Ngày nay các nguồn thông tin thu thập được trong hệ thống gợi ý rất đa dạng, một số thông tin thu được nằm ngoài phạm vi ba loại thông tin nêu trên, ví dụ dữ liệu về mối quan hệ kết bạn (friends) [25], mức độ tin cậy của mối quan hệ từ mạng xã hội (trust) [33], hoặc thông tin đặc trưng có liên quan với nhau của các sản phẩm mà người dùng hay nhóm người dùng có thể thích [28]. Chính vì vậy mà nhu cầu về tính cá nhân hóa với dữ liệu sẽ cao hơn bao giờ hết. Đứng trước thách thức về tính cá nhân hóa dữ liệu tới người dùng lớn như vậy dẫn tới nảy sinh nhu cầu cấp thiết về việc thiết kế, cải tiến và mở rộng các hệ thống gợi ý truyền thống nhằm khai thác tối đa các thông tin khác liên quan giữa người dùng với sản phẩm, từ đó đưa ra gợi ý các sản phẩm, dịch vụ phù hợp nhất với người dùng hiện thời.

Trên cơ sở phân tích những vấn đề còn tồn tại của hai xu hướng chính trong nghiên

cứu về hệ thống gợi ý hiện nay là: Cải tiến các phương pháp lọc tin truyền thống trong hệ thống gợi ý; Mở rộng các phương pháp gợi ý truyền thống cho phép tích hợp thêm các nguồn thông tin khác, tác giả tập trung chính vào nghiên cứu đưa ra hai đề xuất giải quyết bài toán dự đoán năng lực học tập của sinh viên theo hai xu hướng này, các phân tích và kết quả được trình bày cụ thể trong các chương 3,4,5 của luận án.

2.1.3 Các nghiên cứu về dự đoán năng lực học tập của sinh viên

Trong lĩnh vực giáo dục ở Việt Nam, cũng đã có một số nghiên cứu áp dụng khai thác dữ liệu, chủ yếu là sử dụng thông tin cá nhân của sinh viên kết hợp với dữ liệu tuyển sinh để dự đoán kết quả học tập. Phương pháp này thường dựa trên khai thác dữ liệu từ kết quả học tập của các sinh viên trong quá khứ để đưa ra dự báo cho sinh viên đang học. Ngoài ra, còn sử dụng để hỗ trợ cho hoạt động tư vấn học tập, giúp cố vấn học tập tăng cường vai trò của mình trong các hoạt động đào tạo.

Để cung cấp những lời khuyên tư vấn hợp lý cho sinh viên, nhiều nghiên cứu đã tìm cách đánh giá năng lực học tập của họ bằng các phương pháp kỹ thuật phổ biến như mô hình mạng Bayes và cây quyết định [34] [35], luật kết hợp và luật kết hợp dựa trên chuỗi [36] [37], giải thuật di truyền, lý thuyết trò chơi [38], lý thuyết tập thô [39], và thậm chí cả mô hình học sâu [32] [12] hoặc phương pháp lập luận theo tình huống (Case-based reasoning - CBR) [40]. Bảng 2.1 tóm tắt các nghiên cứu này.

Bảng 2.1: Các giải thuật dự đoán năng lực học tập sinh viên theo hướng dự đoán không cá nhân hóa

Giải thuật	Tập dữ liệu	Dữ liệu đầu vào	Dự đoán
Cây quyết định và Mạng Bayes [34]	LNHSGO- (Philippines)	Thông tin nhân khẩu và điểm trung bình học kỳ	Kết quả buộc thôi học
Luật kết hợp và luật kết hợp chuỗi [36]	IUH-2010-2017	Điểm của danh sách các môn đã học	Danh sách các môn học sẽ học hiệu quả

Giải thuật	Tập dữ liệu	Dữ liệu đầu vào	Dự đoán
Học sâu với mạng nơ-ron đa tầng [32]	CTU-2007-2019	Thông tin nhân khẩu, Điểm trung bình học kỳ và điểm ngoại ngữ	Điểm trung bình và điểm môn học
Giải thuật di truyền lai với các giải thuật phân lớp [38]	UAPAZ-Mexico	Thông tin nhân khẩu, thông tin khảo sát và các điểm môn học	Kết quả buộc thôi học
Lý thuyết tập thô [39]	WLAS (Thổ Nhĩ Kỳ)	Thông tin nhân khẩu và điểm trung bình học kỳ trước	Cảnh báo học vụ

Mặc dù đã có nhiều công trình nghiên cứu về dự đoán năng lực học tập sinh viên, tuy nhiên các nghiên cứu này vẫn áp dụng các kỹ thuật truyền thống và cơ sở. Các kết quả dự đoán này vẫn còn có thể cải thiện bởi các nghiên cứu cải tiến sau này.

2.1.4 Các nghiên cứu về cải tiến các mô hình dự đoán

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 [41], luật kết hợp tuần tự [36], giải thuật di truyền [38] hoặc sử dụng lý thuyết tập thô [39] và nhiều nghiên cứu về khai phá dữ liệu giáo dục được trình bày trong tài liệu [42].

Việc tích hợp kỹ thuật học sâu (Deep Learning) vào phân rã ma trận (Matrix Factorization) ngày càng được nhiều nhà nghiên cứu quan tâm bởi sự ưu việt của chúng. Chẳng hạn nghiên cứu [43] đã đề xuất các phương pháp tích hợp học sâu trong lọc cộng tác của hệ thống gợi ý, tuy nhiên những nghiên cứu này chỉ dừng lại ở thực nghiệm ở một vài cách thức tích hợp và so sánh đánh giá giải thuật. Nghiên cứu [44] đã đề xuất kỹ thuật phân rã ma trận sâu (Deep Matrix Factorization) bằng cách ứng dụng kỹ thuật học sâu vào trên các đối tượng người học và tài nguyên học tập trước khi thực hiện giải thuật phân rã ma trận. Cải tiến mô hình bằng việc tích hợp hai mô hình riêng lẻ đang đang

được nhiều nhà nghiên cứu quan tâm, đây cũng là một trong các hướng cải tiến được đề xuất trong luận án.

2.1.5 Các nghiên cứu về tích hợp mối quan hệ dữ liệu

Tuy có nhiều công trình nghiên cứu về RS trong giáo dục, nhưng phần lớn các nghiên cứu này chỉ tập trung khai thác những thuật toán cơ sở [6], do đó các nghiên cứu còn bỏ ngỏ một số vấn đề như mối quan hệ giữa người dùng (user) và thông tin bổ sung cho mục tin (item).

Sự bùng nổ của mạng xã hội trực tuyến đã thúc đẩy nhu cầu tích hợp chúng vào RS, thu hút sự quan tâm của nhiều nhà nghiên cứu. Đã có nhiều nghiên cứu tăng cường độ chính xác của phương pháp MF bằng cách sử dụng mối quan hệ xã hội giữa người dùng [45]. Thay vì chỉ dựa vào việc tìm người dùng có sự tương đồng (similar user), việc kết hợp với các mạng xã hội như Facebook, Twitter và Zalo đã cung cấp thông tin bổ sung cho RS, giúp nâng cao khả năng dự đoán. [33] cung cấp một cái nhìn so sánh về cách tích hợp mạng xã hội vào MF và xác định phương pháp hiệu quả nhất. Tuy nhiên, một trong những hạn chế của giải thuật này là sự phụ thuộc vào mối quan hệ có sẵn trong dữ liệu. Thông qua thực nghiệm, các nghiên cứu đã chứng minh rằng việc kết hợp mối quan hệ mới (độc lập với dữ liệu sẵn có) đã tối ưu hóa hiệu suất của thuật toán.

Một số nghiên cứu khác đã tìm kiếm mối liên quan giữa các mục tin (item) để tích hợp vào MF nhằm nâng cao hiệu quả cho giải thuật. [46] đã sử dụng lại việc tìm những mục tin có đánh giá gần giống nhau (similar item) để nâng cao hiệu quả dự đoán, với nghiên cứu này phải phụ thuộc vào việc người dùng có đánh giá các mục tin hay không. Nghiên cứu [8] đã kết hợp các thuộc tính vào mục tin để áp dụng kỹ thuật MF, cải tiến này phù hợp cho dữ liệu mục tin có nhiều thông tin và bỏ qua sự liên quan giữa các mục tin. Tuy nhiên những nghiên cứu này ứng dụng ở các lĩnh vực khác, cũng như phương pháp khác. Việc bổ sung mối quan hệ các mục tin với nhau vào kỹ thuật MF cũng là hướng nghiên cứu của luận án.

2.2 Tổng quan về hệ thống gợi ý

2.2.1 Giới thiệu về hệ thống gợi ý

Hệ thống gợi ý (Recommender Systems - RS) là một loại hệ thống máy tính được thiết kế để đưa ra các gợi ý về sản phẩm, dịch vụ hoặc thông tin mà người dùng có thể quan tâm dựa trên dữ liệu lịch sử của họ. Các hệ thống gợi ý được sử dụng rộng rãi trong nhiều lĩnh vực, bao gồm thương mại điện tử, truyền thông xã hội, bài viết, âm nhạc, phim ảnh, sách báo và nhiều lĩnh vực khác. Hệ thống gợi ý hoạt động bằng cách thu thập dữ liệu về hành vi của người dùng, sau đó sử dụng các thuật toán máy học và trí tuệ nhân tạo để phân tích dữ liệu này và tạo ra các gợi ý dựa trên thông tin đó.

Hiện nay, với sự phát triển của internet và nhiều dịch vụ trực tuyến, người dùng có rất nhiều thông tin để xem xét trước khi quyết định mua sản phẩm hoặc sử dụng dịch vụ điều này có thể khiến cho người dùng cảm thấy quá tải và không biết nên chọn lựa sản phẩm hoặc dịch vụ nào là tốt nhất cho mình, dẫn đến việc lựa chọn không đúng có thể dẫn đến sự mất mát của thời gian và tiền bạc. Hệ thống gợi ý [47] là một công cụ phần mềm tích hợp vào các ứng dụng hoặc trang web, nhằm giảm bớt sự quá tải thông tin cho người dùng. Hệ thống này sử dụng các loại dữ liệu và tri thức khác nhau (như sở thích, hành động và ngữ cảnh của người dùng) để đề xuất các mục hữu ích cho người dùng, như sản phẩm, bộ phim hay bài hát. Mục đích của hệ thống gợi ý là giúp người dùng tiết kiệm thời gian và nỗ lực trong việc tìm kiếm thông tin, đồng thời nâng cao trải nghiệm của họ khi sử dụng các ứng dụng hoặc trang web.

Trong hai thập kỷ qua, người ta đã chứng kiến một sự quan tâm ngày càng tăng về các hệ thống gợi ý [30]. RS đã trở thành một công cụ quan trọng trong việc giúp con người xử lý thông tin trên internet, đặc biệt là trong lĩnh vực thương mại điện tử. Tuy nhiên, các nhà nghiên cứu vẫn đang tiếp tục phát triển và nghiên cứu các vấn đề liên quan đến RS để cải thiện chất lượng và hiệu quả của các hệ thống này. Bên cạnh đó, với sự phát triển đa dạng của ứng dụng RS trong thực tế, nhiều ngành nghề khác cũng đang quan tâm và áp dụng RS để giúp người dùng đưa ra các quyết định chính xác và đạt được sự hài lòng tốt nhất.

Hệ thống gợi ý có thể hỗ trợ người dùng tìm kiếm và khám phá các tài nguyên khác

nhau, bao gồm phim ảnh, âm nhạc, sách, các trang web, tin tức trực tuyến, địa điểm du lịch và thậm chí là phong cách sống. Hệ thống này hoạt động dựa trên thông tin cá nhân và đánh giá phản hồi từ người dùng. Đánh giá phản hồi có thể thu thập thông qua các hình thức tường minh hoặc không tường minh, như là việc đánh giá sản phẩm hoặc sự tương tác của người dùng với hệ thống, bao gồm các hành động như click chuột, thời gian quan sát và đặt hàng. Vì vậy, hệ thống gợi ý là một công cụ hữu ích để giảm tình trạng quá tải thông tin và cung cấp các lời khuyên, đề xuất phù hợp với sở thích và nhu cầu của người dùng. Nó đã được áp dụng rộng rãi trong nhiều lĩnh vực, đặc biệt là trong thương mại điện tử, giúp các khách hàng tìm kiếm và mua sản phẩm dễ dàng hơn. Tuy nhiên, vẫn còn rất nhiều vấn đề cần nghiên cứu để phát triển và cải tiến hệ thống gợi ý để đáp ứng nhu cầu ngày càng đa dạng của người dùng.

2.2.2 Các lĩnh vực ứng dụng của hệ thống gợi ý

Hiện tại hệ thống gợi ý được ứng dụng trong nhiều lĩnh vực khác nhau và được các nghiên cứu [48] [5] tổng hợp như sau:

- Trong thương mại điện tử có gợi ý sản phẩm và dịch vụ, ví dụ như hệ thống gợi ý của Amazon, Alibaba.
- Trong giáo dục có gợi ý các nguồn tài nguyên cho học tập như sách tham khảo, bài báo khoa học, địa chỉ web cho người học, ví dụ hệ thống InfoFinder, Foxtrot.
- Trong giải trí có gợi ý bài hát, hoặc phim ảnh đến người dùng (như hệ thống LastFM, Netflix, MovieLens, YouTube).
- Trong du lịch có gợi ý địa điểm tham quan, hoạt động du lịch cho khách tham quan (như hệ thống DieToRecs, Lifestyle Finder).
- Trong y tế có gợi ý các sản phẩm y tế, như hệ thống PatientsLikeMe.
- Trong mạng xã hội trực tuyến có gợi ý các bài viết, kết bạn, và các hoạt động xã hội khác như hệ thống Facebook, Zalo.
- Trong kinh doanh các dịch vụ ăn uống có gợi ý các địa điểm nhà hàng, quán ăn, món ăn ví dụ hệ thống Adaptive Place Advisor, Akeedapp, Pocket Restaurant Finder.

Bên cạnh đó, các hệ thống gợi ý đã và đang được các nhà nghiên cứu, các tổ chức và doanh nghiệp rất quan tâm, để ứng dụng hệ thống gợi ý cho đa dạng các lĩnh vực trong cuộc sống.

Thật vậy, trong thế giới số hóa ngày nay, việc học tập trực tuyến đã trở thành một xu hướng không thể phủ nhận, giúp hàng triệu người trên khắp thế giới tiếp cận kiến thức một cách dễ dàng và linh hoạt. Nhờ vào sự phổ biến của các khóa học trực tuyến, nhu cầu về việc tạo ra một trải nghiệm học tập cá nhân hóa và hiệu quả hơn cũng ngày càng tăng. Đó chính là lý do Hệ trợ giảng thông minh (Intelligent Tutoring Systems - ITS) được nhiều nhà nghiên cứu quan tâm. Những hệ thống này không chỉ mang lại kiến thức mà còn phân tích, đánh giá và tối ưu hóa quá trình học tập dựa trên từng cá nhân, mở ra một kỷ nguyên mới cho giáo dục trực tuyến. Phần tiếp theo luận án sẽ trình bày chi tiết hơn về hệ thống trợ giảng thông minh.

2.2.3 Giới thiệu hệ thống trợ giảng thông minh

Hệ trợ giảng thông minh (ITS) là một hệ thống tiên tiến, có khả năng tự động điều chỉnh nội dung và cung cấp phản hồi ngay lập tức cho học viên, mà không yêu cầu sự hỗ trợ trực tiếp từ giáo viên. Đặc điểm nổi bật của ITS là sự kết hợp giữa ba lĩnh vực quan trọng: Khoa học máy tính (cung cấp công nghệ và thuật toán), giáo dục học (định hình phương pháp và chiến lược giảng dạy) và tâm lý học (hiểu về quá trình học và cảm nhận của học viên). Mặc dù từng hệ thống ITS có thể được thiết kế riêng biệt và có những yếu tố khác nhau tùy theo mục tiêu và lĩnh vực ứng dụng, nhưng có bốn thành phần chính thường xuất hiện trong hầu hết các ITS: (1) Mô hình người học (Student Model), (2) Mô hình tri thức (Domain/Knowledge Model), (3) Mô hình trợ giảng (Instructor/Tutoring Model) và (4) Giao diện người dùng (User Interface).

Mô hình tri thức (Domain Model) chứa tập hợp các kiến thức, kỹ năng, thái độ và kế hoạch giảng dạy của lĩnh vực học tập [49]. Kiến thức thường là những kiến thức chuyên môn chủ yếu của mỗi lĩnh vực học tập. Kỹ năng là những nguyên tắc, ràng buộc thực hành, kể cả những lỗi thường mắc phải của người học. Thái độ là những quan niệm tích cực hoặc sai lầm mà người học thể hiện định kỳ xuyên suốt trong quá trình học.

Mô hình trợ giảng (Tutoring Model), còn được gọi là mô hình sư phạm hoặc mô hình hướng dẫn, lấy mô hình tri thức và mô hình người học làm đầu vào và lựa chọn các hành động trợ giảng phù hợp để phản hồi tức thời đến người học [50]. Trong các hệ thống ITS, ví dụ như [51], người học cũng có thể thực hiện các hành động như đặt câu hỏi hoặc yêu cầu trợ giúp thì hệ thống luôn sẵn sàng phản hồi và quyết định những hành động tiếp

theo tại bất kỳ thời điểm nào trong quá trình dạy và học. Những phản hồi này được xác định bởi một mô hình trợ giảng tuân thủ các quy định và lý thuyết sư phạm mà các nhà khoa học đã nghiên cứu.

Giao diện người dùng (User Interface) là giao diện người-máy tính hỗ trợ người học tương tác hệ thống thông qua các phương tiện đầu vào khác nhau (giọng nói, gõ phím, nhấp chuột, video quan sát) và tạo ra đầu ra ở các phương tiện khác nhau (văn bản, sơ đồ, hoạt ảnh, tác nhân). Ngoài các tính năng giao diện người-máy tính thông thường, một số hệ thống gần đây đã kết hợp tương tác ngôn ngữ tự nhiên [52], nhận dạng giọng nói [53], và cảm nhận cảm xúc của người học [54].

Mô hình người học (Student Model) là mô hình đánh giá năng lực của người học bao gồm các năng lực nhận thức, tình cảm, động cơ và các trạng thái tâm lý khác của người học được phát triển trong quá trình học [55]. Nó thường được xem như tập hợp con của mô hình tri thức (tập hợp tri thức mà người học đã đạt được), mô hình này sẽ thay đổi trong quá trình dạy và học. Nhiệm vụ quan trọng của mô hình này là đánh giá năng lực học tập, là theo dõi sự tiến bộ của người học từ vấn đề này sang vấn đề khác, cũng như xây dựng hồ sơ năng lực của người học (điểm mạnh, điểm yếu, kết quả học tập). Ngoài việc đánh giá năng lực học tập, mô hình người học còn dự đoán năng lực học tập của người học để điều chỉnh các hướng dẫn trong mô hình trợ giảng.

Tóm lại, các phần trên là một số yếu tố cần được xem xét và thách thức khi xây dựng mô hình người học trong ITS. Trong các thành phần chính của ITS thì thành phần quan trọng nhất là mô hình người học. Xét chi tiết hơn, trong mô hình người học thì vấn đề dự đoán năng lực người học giữ vai trò trung tâm và quan trọng nhất. Chính vì thế trong khuôn khổ luận án không thể nghiên cứu tất cả thành phần của ITS hoàn chỉnh, mà chỉ tập trung nghiên cứu thành phần quan trọng nhất và có khả năng tận dụng được thành tựu của lĩnh vực khoa học máy tính, đó là bài toán dự đoán năng lực học tập khi người học đang giải quyết vấn đề (bài tập) cụ thể nào đó [55] để đưa ra gợi ý những môn học cho người học có thể học tốt và phù hợp với năng lực.

2.2.4 Các nguồn tài nguyên về hệ thống gợi ý

Hệ thống gợi ý không chỉ đơn thuần là một dạng hệ thống thông tin, mà nó còn là một lĩnh vực nghiên cứu đang nhận được sự quan tâm lớn từ các nhà khoa học. Để hỗ

trợ cho việc nghiên cứu và phát triển các giải thuật của hệ thống gợi ý, nhiều tập dữ liệu như LastFM, MovieLens và cũng như các phần mềm, thư viện nguồn mở đã được cung cấp. Các công cụ này giúp tăng tốc độ phát triển sản phẩm và nghiên cứu hệ thống gợi ý. Bảng 2.2 dưới đây giới thiệu một số mã nguồn mở được cộng đồng cung cấp để hỗ trợ việc xây dựng ứng dụng hệ thống gợi ý [29].

Bảng 2.2: Một số mã nguồn mở để phát triển hệ thống gợi ý

Mã nguồn	Miêu tả	Ngôn ngữ
Apache Mahout	Thư viện máy học nhằm xử lý dữ liệu lớn, trong đó có các thuật toán tiên tiến về lọc cộng tác. Việc tích hợp thư viện này khá đơn giản. www.mahout.apache.org	Java
Cofi	Thư viện các thuật toán lọc cộng tác ở dạng cơ bản. www.nongnu.org/cofi	Java
Crab	Thư viện đa dạng các thành phần giúp xây dựng và tùy chỉnh hệ thống gợi ý từ các thuật toán truyền thống. www.github.com/muricoca/crab	Python
Easyrec	Framework dùng để xây dựng hệ thống gợi ý trên nền Web. www.easyrec.org	Java
LensKit	Thư viện lọc cộng tác cơ sở được phát triển bởi nhóm nghiên cứu GroupLens. www.lenskit.grouplens.org	Java
My Media Lite	Thư viện cài đặt các giải thuật, nhóm giải thuật phân rã ma trận, nhằm hỗ trợ xây dựng và kiểm nghiệm hệ thống gợi ý bằng ngôn ngữ C#. www.mymedialite.net	C#
PREA	Bộ công cụ cài đặt các thuật toán gợi ý. www.prea.gatech.edu	Java

Mã nguồn	Miêu tả	Ngôn ngữ
SVD Feature	Thư viện cung cấp các cài đặt các giải thuật thuộc kỹ thuật Matrix Factorization. www.mloss.org/software/view/333/	C++
Vogoo	Thư viện các phương pháp lọc cộng tác cho xây dựng hệ thống gợi ý trên nền Web. www.sourceforge.net/projects/vogoo/	PHP
MLib	Thư viện các phương pháp lọc cộng tác dựa trên mô hình, sử dụng thuật toán ALS (Alternating Least Squares) để cài đặt. spark.apache.org/docs/latest/ml-collaborative-filtering.html	Java
LibRec	Thư viện cung cấp đa dạng các cài đặt hệ thống gợi ý. Thư viện này có sẵn các thuật toán gợi ý từ cơ bản đến nâng cao, gồm các thuật toán tiên tiến hiện nay (hỗ trợ hơn 70 thuật toán). www.librec.net	Java
Recom-mender Lab	Thư viện cung cấp các hàm cơ bản để xây dựng và kiểm thử các thuật toán gợi ý trong một framework tiêu chuẩn, từ đó giúp xử lý và phân tích dữ liệu mới đưa vào một cách nhanh chóng. www.github.com/mhahsler/recommenderlab	R

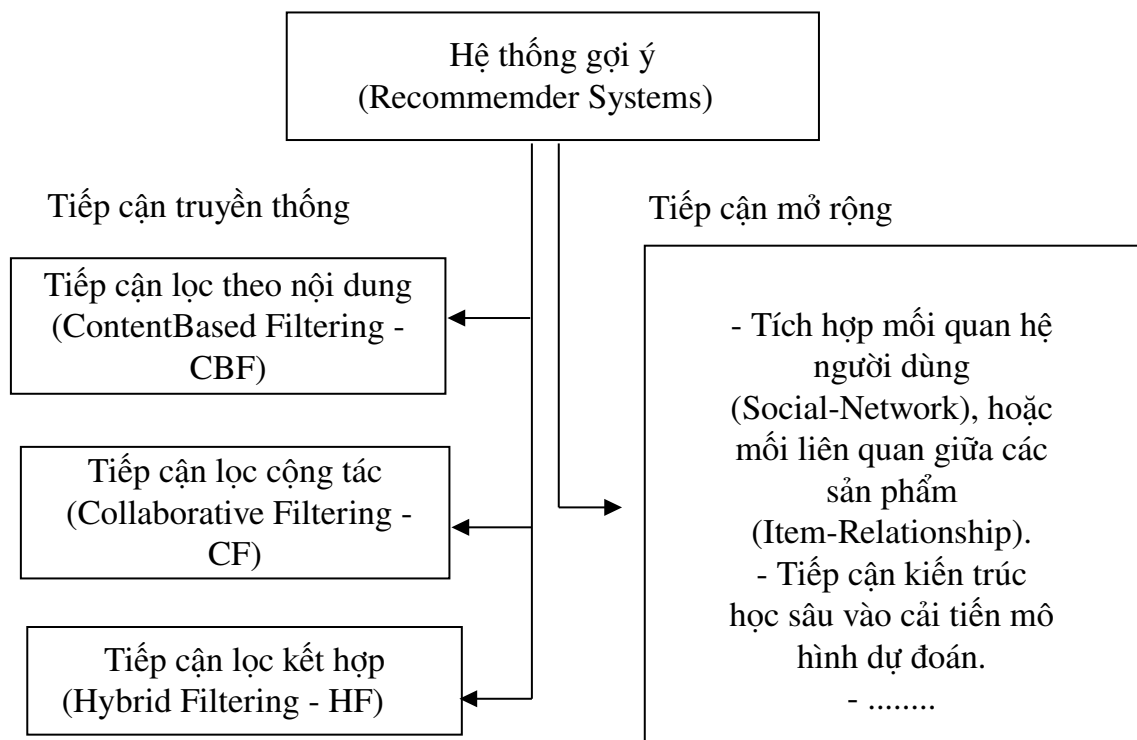
2.2.5 Các hướng tiếp cận để xây dựng hệ thống gợi ý

Có nhiều nghiên cứu khác nhau để giải quyết bài toán gợi ý theo “Quy trình xây dựng hệ thống gợi ý”. Tuy nhiên về cơ bản thì hệ thống gợi ý được chia thành hai hướng tiếp cận tùy vào việc lựa chọn loại thông tin, mô hình học và dự đoán sản phẩm mới cho người dùng [56] được trình bày trong Hình 2.1, đó là: (1) Hệ thống gợi ý với cách tiếp cận truyền thống; (2) Hệ thống gợi ý mở rộng. Trong đó:

- Cách tiếp cận truyền thống khai thác ba loại thông tin đầu vào gồm người dùng, mục tin và phản hồi của người dùng về mục tin. Dựa vào cách xác định dự đoán đánh giá cho các mục tin đối với người dùng, hệ thống gợi ý thường được chia thành ba loại: Gợi ý dựa vào phương pháp lọc cộng tác (Collaborative Filtering), gợi ý dựa vào

phương pháp lọc theo nội dung (Content-Based Filtering) và gợi ý dựa vào phương pháp lọc kết hợp (Hybrid Filtering) [57].

- Cách tiếp cận mở rộng từ hệ thống gợi ý truyền thống cho phép tích hợp thêm các nguồn thông tin khác (thông tin trong mạng xã hội, mối quan hệ giữa các mục đối tượng) hoặc cải tiến các phương pháp lọc tin truyền thống trong hệ thống gợi ý (các phương pháp lọc kết hợp, các phương pháp lọc dựa trên mối quan tâm, hoặc tích hợp các kỹ thuật học sâu vào các kỹ thuật truyền thống). Từ đây hệ thống gợi ý được chia thành một số loại điển hình: Gợi ý dựa vào ngữ cảnh (Context-Aware) [58], gợi ý dựa vào mạng xã hội (Social Network-Based) [25], gợi ý dựa vào các phương pháp lọc kết hợp (Hybrid Filtering) [59], tiếp cận các phương pháp lọc dựa trên mối quan tâm (Attention-based) [26], tiếp cận tích hợp các kỹ thuật học sâu (Deep Learning) [31] .



Hình 2.1: Các hướng tiếp cận truyền thống và xu hướng hiện nay của hệ thống gợi ý

2.2.5.1 Lọc nội dung

Phương pháp lọc theo nội dung: Sử dụng các thuật toán lọc thông tin để loại bỏ các tin tức không liên quan và độc đáo và sử dụng các thuật toán lọc nội dung để đề xuất các tin tức có chủ đề tương tự nhưng không giống hệt nhau. Phương pháp lọc nội dung có

tính linh hoạt cao vì chúng có thể được tùy chỉnh cho từng người dùng hoặc nhóm người dùng cụ thể, tùy thuộc vào sở thích và nhu cầu của họ, có tính khả dụng có thể hoạt động với bất kỳ loại nội dung nào, từ văn bản đến video và âm nhạc, giúp người dùng tìm thấy nội dung phù hợp với mình và đưa ra các đề xuất nội dung chính xác và phù hợp, tăng cường trải nghiệm của người dùng và giúp họ tiết kiệm thời gian tìm kiếm nội dung.

Ngoài ra, phương pháp này cũng còn vài một hạn chế như:

- Giới hạn thông tin có thể gây ra hiện tượng lọc thông tin, chỉ đưa ra những đề xuất nội dung giống như những gì người dùng đã tiêu thụ trước đó. Điều này có thể dẫn đến việc người dùng bỏ lỡ các nội dung mới và khám phá.
- Khó đo lường hiệu quả: Đánh giá hiệu quả của các phương pháp lọc nội dung là một thách thức, vì độ chính xác của các đề xuất phụ thuộc vào nhiều yếu tố khác nhau, bao gồm sự phù hợp với nhu cầu và sở thích của từng người dùng cũng như sự đa dạng của nội dung.
- Yêu cầu dữ liệu đầu vào chất lượng cao: Các phương pháp lọc nội dung yêu cầu dữ liệu đầu vào chất lượng cao để đưa ra các đề xuất nội dung chính xác. Nếu dữ liệu đầu vào không đủ hoặc không chính xác, các đề xuất nội dung có thể không phù hợp với sở thích và nhu cầu của người dùng.

Ý tưởng của hệ thống gợi ý sử dụng phương pháp lọc theo nội dung là gợi ý một tài nguyên đến người dùng dựa trên việc so sánh độ tương đồng giữa nội dung tài nguyên và các đặc trưng của người dùng, những tài nguyên có độ tương đồng cao sẽ được chọn để gợi ý [27]. Ví dụ, một người thích phim khoa học viễn tưởng thì những phim có nội dung liên quan đến khoa học viễn tưởng sẽ được gợi ý đến người dùng đó [56]. Các phương pháp dựa trên tiếp cận nội dung thông thường sẽ thực hiện các bước sau:

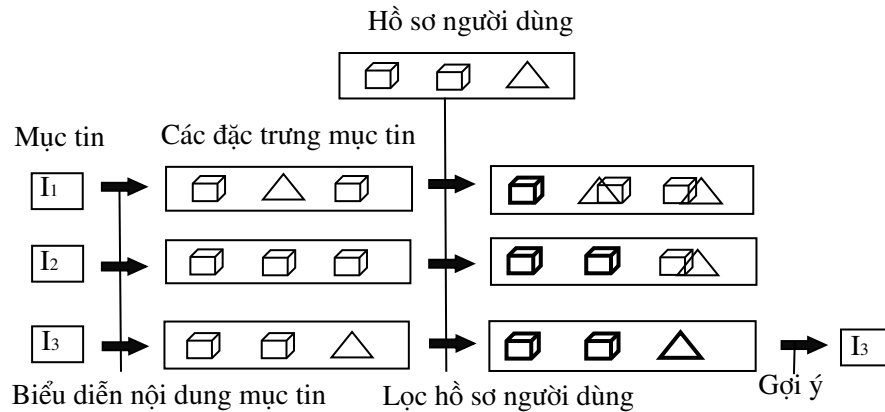
- Bước 1. Biểu diễn nội dung của đối tượng khuyến nghị $i \in I$, được ký hiệu là $ContentItem(i)$, thông qua tập $|C|$ đặc trưng nội dung của i . Tập các đặc trưng của mục tin i được xây dựng bằng các kỹ thuật truy vấn thông tin.

- Bước 2. Mô hình hóa sở thích người dùng $u \in U$, gọi tắt là hồ sơ người dùng, ký hiệu $UserProfile(u)$. Hồ sơ của người dùng thực chất là lịch sử truy cập hoặc đánh giá của người đó đối với các đặc trưng nội dung mục tin. $UserProfile(u)$ được xây dựng bằng cách phân tích nội dung các mục tin mà người dùng u đã từng truy nhập hoặc đánh giá dựa trên các kỹ thuật truy vấn thông tin. Do vậy $UserProfile(u)$ là một véc-tơ biểu diễn

thông qua $|C|$ đặc trưng nội dung của I .

- Bước 3. Dự đoán đánh giá của người dùng u với mục tin i dựa trên độ tương tự nội dung của i với hồ sơ người dùng u . Hệ thống sẽ ưu tiên gợi ý những đối tượng i có nội dung tương tự cao nhất với hồ sơ người dùng u .

Tiến trình xử lý của hệ thống gợi ý sử dụng phương pháp lọc theo nội dung được cụ thể hóa trong Hình 2.2 như sau:



Hình 2.2: Tiến trình xử lý của hệ thống gợi ý lọc theo nội dung

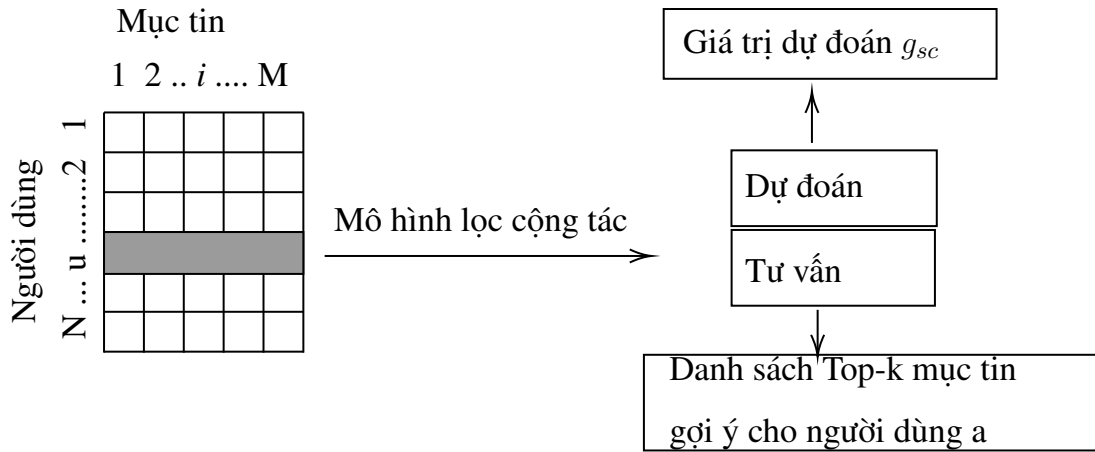
Các phương pháp gợi ý truyền thống dựa trên nội dung có thể chia thành hai nhóm chính: (1) Lọc theo nội dung dựa vào bộ nhớ, thực hiện tính toán độ tương tự giữa $ContentItem(i)$ và vẩy $UserProfile(u)$ dùng các độ đo tương tự như Cosine, Euclidean... để thực hiện huấn luyện và dự đoán; (2) Lọc theo nội dung dựa vào mô hình, với mô hình được học từ dữ liệu dùng các kỹ thuật thống kê hoặc học máy như: mạng Bayes, phân cụm, cây quyết định, mạng nơ-ron nhân tạo... [21] để phân các đối tượng khuyến nghị thành những đối tượng người dùng quan tâm hay không quan tâm.

2.2.5.2 Lọc cộng tác

Hệ thống lọc cộng tác giới thiệu mục tin phù hợp cho một người dùng dựa trên sở thích của những người dùng có đặc điểm tương đồng với người dùng đó. Nếu hầu hết những người dùng tương đồng với người dùng u đều thích mục tin i , thì mục tin i cũng sẽ được gợi ý cho người dùng u . Tập hợp những người dùng tương đồng nhau sẽ hình thành một cộng đồng.

Phương pháp lọc cộng tác khai thác thông tin liên quan đến hành vi đối với mục tin của cộng đồng người dùng có sở thích tương đồng trong quá khứ để dự đoán các mục tin

mới phù hợp với người dùng hiện tại [5]. Do đó, thông tin đầu vào cho hệ thống gợi ý dựa trên lọc cộng tác là các phản hồi từ người dùng về các mục tin trong hệ thống, được biểu diễn bằng ma trận đánh giá. Quy trình xây dựng hệ thống gợi ý sử dụng phương pháp lọc cộng tác được trình bày chi tiết trong (Hình 2.3).



Hình 2.3: Tiến trình xử lý của hệ tư vấn sử dụng lọc cộng tác

Phương pháp lọc cộng tác đã được xác nhận là một trong những chiến lược thành công nhất khi xây dựng hệ thống gợi ý. Những đặc tính như sự tiện lợi trong quá trình triển khai và khả năng xử lý nhiều dạng thông tin - từ văn bản đến đa phương tiện như hình ảnh và âm thanh - làm nổi bật nó so với các phương pháp khác. Rất nhiều công trình nghiên cứu đã dành thời gian để khảo sát, tổ chức và đánh giá các thuật toán dựa trên lọc cộng tác. Mỗi kỹ thuật mang lại những lợi ích và thách thức riêng, đều tập trung vào việc tìm hiểu các liên kết trong ma trận đánh giá của người dùng. Đại khái, các kỹ thuật lọc cộng tác có thể được phân thành hai danh mục chính.

- Lọc cộng tác dựa vào bộ nhớ (hoặc Heuristic-based) giới thiệu nhiều chiến lược tiếp cận khác nhau [5]. Đầu tiên, chúng ta có gợi ý dựa trên người dùng (User-Based Collaborative Filtering) và gợi ý dựa trên mục tin (Item-Based Collaborative Filtering). Tiếp theo, chúng ta cũng xem xét việc áp dụng trọng số tương tự (Significance Weighting), thiết lập giá trị mặc định cho các đánh giá (Default Voting), và chuẩn hóa giá trị trong ma trận đánh giá thông qua tần suất xuất hiện ngược của người dùng (Inverse User Frequency). Cuối cùng, chúng ta có kỹ thuật tăng cường mối liên giữa những người dùng gần nhau (Case Amplification). Mặc dù mỗi hướng tiếp cận tương tác với ma trận đánh giá theo cách riêng của nó để tối ưu hóa việc đề

xuất, hai phương pháp đứng đầu về sự phổ biến là gợi ý dựa trên người dùng và dựa trên mục tin. Chi tiết về hai kỹ thuật này sẽ được thảo luận trong phần tiếp theo.

- Phương pháp lọc cộng tác dựa vào mô hình (Model-based) tạo ra một khác biệt rõ ràng so với lọc cộng tác dựa vào bộ nhớ. Theo [21], thay vì chỉ sử dụng ma trận đánh giá trực tiếp, phương pháp này tập trung vào việc xây dựng mô hình từ ma trận đó để dự đoán và đề xuất sản phẩm cho người dùng. Lợi ích chính của việc sử dụng mô hình là kích thước mô hình thường nhỏ hơn nhiều so với ma trận đánh giá gốc và có khả năng dự đoán nhanh hơn. Chỉ khi có sự thay đổi lớn trong dữ liệu, mô hình mới cần được huấn luyện lại. Đặc biệt, trong lọc cộng tác dựa vào mô hình, chúng ta có thể áp dụng nhiều kỹ thuật học máy và khai thác dữ liệu như luật kết hợp [31], phân cụm [60], SVM [61], cây quyết định [62], mạng nơ-ron nhân tạo [43], học sâu [63], đồ thị, phân loại, hồi qui, mạng Bayes [64], và phân rã ma trận. Các phương pháp sử dụng Ontology và giảm chiều dữ liệu cũng được áp dụng [65] [66] để tối ưu hóa mô hình dự đoán.

Ngoài những tính hay của mô hình lọc cộng tác cũng có một số nhược điểm:

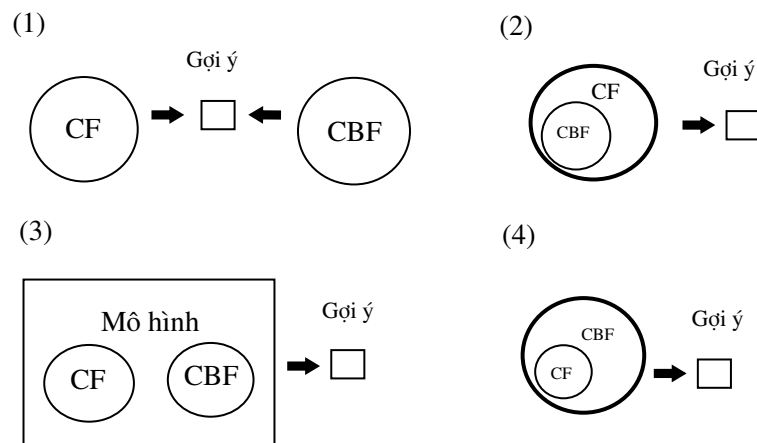
- Mô hình lọc cộng tác cần một lượng lớn dữ liệu về hành vi của người dùng để có thể tạo ra các gợi ý chính xác. Khi hệ thống mới ra đời hoặc có ít dữ liệu, mô hình lọc cộng tác sẽ không thể hoạt động hiệu quả và cung cấp các gợi ý không chính xác. Vấn đề hiệu quả khi số lượng sản phẩm quá lớn: Với số lượng sản phẩm lớn, việc tính toán các độ tương đồng giữa các sản phẩm sẽ tốn nhiều thời gian và tài nguyên tính toán. Do đó, mô hình lọc cộng tác có thể trở nên chậm và không thể tạo ra các gợi ý kịp thời.
- Vấn đề đa dạng của các gợi ý: Mô hình lọc cộng tác có thể tạo ra các gợi ý giống nhau dẫn đến sự thiếu đa dạng và khó khăn trong việc đáp ứng nhu cầu người dùng.
- Vấn đề chủ quan của đánh giá người dùng: Mô hình lọc cộng tác dựa trên đánh giá của người dùng để đưa ra các gợi ý. Tuy nhiên, đánh giá của người dùng có thể bị thiên vị hoặc không chính xác, dẫn đến các gợi ý không phù hợp với nhu cầu của người dùng.
- Vấn đề về tính riêng tư: Mô hình lọc cộng tác yêu cầu thu thập dữ liệu về hành vi của người dùng để tạo ra các gợi ý. Việc thu thập dữ liệu này có thể gây ra vấn đề về quyền riêng tư của người dùng, đặc biệt là khi các dữ liệu này bị sử dụng mà không

được sự đồng ý của người dùng.

2.2.5.3 Lọc kết hợp

Dựa vào đặc trưng thông tin, chúng ta có thể lọc nội dung dựa trên các khía cạnh liên quan của các đối tượng muốn tìm. Ngược lại, việc lọc cộng tác tập trung vào việc phân tích sở thích cá nhân của người dùng dựa trên việc họ tiếp xúc với loại thông tin nào. Mỗi cách tiếp cận đều có điểm mạnh và điểm yếu riêng. Do đó, để tối ưu hóa lợi ích và giảm thiểu hạn chế của từng kỹ thuật, phương pháp lọc kết hợp đã được ra đời, giúp tăng cường khả năng gợi ý sản phẩm phù hợp cho người dùng [24]. Nghiên cứu đã chỉ ra rằng lọc kết hợp mang lại dự đoán chính xác hơn so với lọc nội dung hoặc lọc cộng tác riêng lẻ. Đặc biệt, nó giảm thiểu vấn đề dữ liệu thưa và người dùng mới vào hệ thống [30].

Lọc kết hợp là sự tổng hợp của nhiều kỹ thuật gợi ý [67] [5]. Bốn hướng tiếp cận chính, như minh họa trong Hình 2.4, bao gồm: 1) Kết hợp dự đoán từ cả lọc cộng tác và lọc nội dung [31]; 2) Tích hợp đặc điểm từ lọc nội dung vào phương pháp lọc cộng tác [62]; 3) Tích hợp đặc điểm từ lọc cộng tác vào phương pháp lọc nội dung [68]; và 4) Phát triển một mô hình tổng hợp giữa lọc cộng tác và lọc nội dung [69].



Hình 2.4: Các phương pháp lọc kết hợp giữa lọc cộng tác và lọc nội dung

Mô hình lọc kết hợp (hybrid filtering) là một phương pháp kết hợp giữa lọc dựa trên nội dung (content-based filtering) và lọc dựa trên hành vi (collaborative filtering) để cải thiện khả năng dự đoán trong hệ thống gợi ý.

Lọc kết hợp cung cấp độ chính xác cao hơn cho hệ thống dự đoán: Khi kết hợp các phương pháp khác nhau, mô hình lọc kết hợp có thể tận dụng tốt những ưu điểm của

từng phương pháp và giảm thiểu những hạn chế của chúng, giúp cải thiện độ chính xác và độ tin cậy của hệ thống dự đoán. Tăng tính đa dạng cho các gợi ý: Mô hình lọc kết hợp có thể cung cấp gợi ý cho người dùng dựa trên các thông tin khác nhau, bao gồm cả nội dung của sản phẩm và thông tin về hành vi của người dùng, giúp tăng tính đa dạng và phù hợp của các gợi ý. Tính linh hoạt: Mô hình lọc kết hợp cho phép thay đổi và tùy chỉnh các thành phần của nó để phù hợp với từng hệ thống gợi ý cụ thể.

Tuy nhiên, mô hình lọc kết hợp cũng có một số hạn chế như:

- **Đòi hỏi nhiều tài nguyên và thời gian để triển khai:** Mô hình lọc kết hợp yêu cầu tính toán phức tạp để tính toán đánh giá và đưa ra các gợi ý chính xác. Do đó, nó đòi hỏi nhiều tài nguyên và thời gian để triển khai và hoạt động.
- **Khó khăn trong việc tính toán đánh giá và phân tích dữ liệu:** Mô hình lọc kết hợp phải tính toán đánh giá cho nhiều loại dữ liệu khác nhau, bao gồm thông tin về sản phẩm, dữ liệu người dùng và dữ liệu hành vi. Điều này có thể gây ra khó khăn trong việc phân tích dữ liệu và tính toán đánh giá độ chính xác của các gợi ý.

2.3 Bài toán dự đoán kết quả học tập của sinh viên

Bài toán dự đoán kết quả học tập của sinh viên có thể rất hữu ích trong nhiều tình huống khác nhau tại các trường đại học. Ví dụ như xác định các ứng viên tiềm năng cho các cuộc thi tài năng, hoặc cấp học bổng nhằm khuyến khích sinh viên nỗ lực hơn trong học tập. Ngoài ra, việc xác định các sinh viên có năng lực yếu (có nguy cơ bị buộc thôi học) cũng rất quan trọng để có thể đưa ra các biện pháp hỗ trợ thích hợp nhằm giúp các sinh viên này học tập tốt hơn. Có hai dạng bài toán dự đoán kết quả học tập của sinh viên: dự đoán cá nhân hóa và dự đoán quy luật chung (không cá nhân hóa). Hai ví dụ cụ thể về hai phương pháp này sẽ được trình bày tiếp theo.

2.3.1 Dự đoán theo cá nhân hóa

Dự đoán cá nhân hóa là một lĩnh vực của học máy và trí tuệ nhân tạo mà trong đó các mô hình và hệ thống được tối ưu hóa để cung cấp dự đoán và trải nghiệm duy nhất cho từng cá nhân dựa trên thông tin cá nhân và lịch sử tương tác của họ. Thay vì mô hình hóa dữ liệu dưới dạng đại diện tổng quát, các giải thuật cá nhân hóa hướng đến việc hiểu rõ

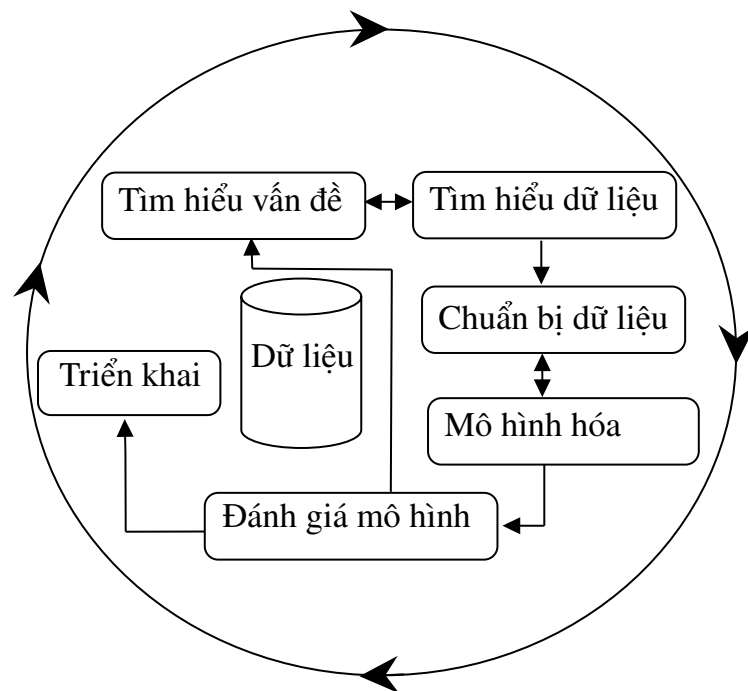
sở thích, nhu cầu và bối cảnh cá nhân của người sử dụng.

Quá trình này thường bắt đầu từ việc thu thập dữ liệu cá nhân, như lịch sử tương tác, thói quen, và thông tin hồ sơ. Các mô hình sau đó sử dụng dữ liệu này để tạo ra dự đoán và gợi ý cá nhân hóa. Việc này có thể áp dụng trong nhiều lĩnh vực, bao gồm gợi ý sản phẩm, nội dung trực tuyến, giáo dục, y tế, và nhiều hơn nữa.

Dự đoán cá nhân hóa không chỉ tăng cường trải nghiệm người dùng mà còn mang lại những lợi ích lớn cho doanh nghiệp và tổ chức. Bằng cách tối ưu hóa sự tương tác và cung cấp thông điệp có ý nghĩa, cá nhân hóa giúp tăng cường sự thân thiện và hiệu suất của hệ thống, tạo ra một môi trường tương tác đặc sắc và tối ưu cho mỗi người dùng.

Bài toán dự đoán kết quả học tập của sinh viên là dạng bài toán dự đoán cá nhân hóa trong lĩnh vực giáo dục. Dựa vào thông tin cá nhân, bảng điểm cá nhân, một số hoạt động, từ đó đưa ra những dự đoán và gợi ý thông tin phù hợp với từng cá nhân cụ thể.

Bài toán dự đoán kết quả học tập của sinh viên thường áp dụng quy trình CRISP-DM (CRoss Industry Standard Process for Data Mining), một tiêu chuẩn trong khai thác dữ liệu. Quy trình này bao gồm sáu bước từ phân tích vấn đề đến triển khai. Hình 2.5 mô tả chi tiết quy trình này.



Hình 2.5: Quy trình khai thác dữ liệu chuẩn CRISP-DM

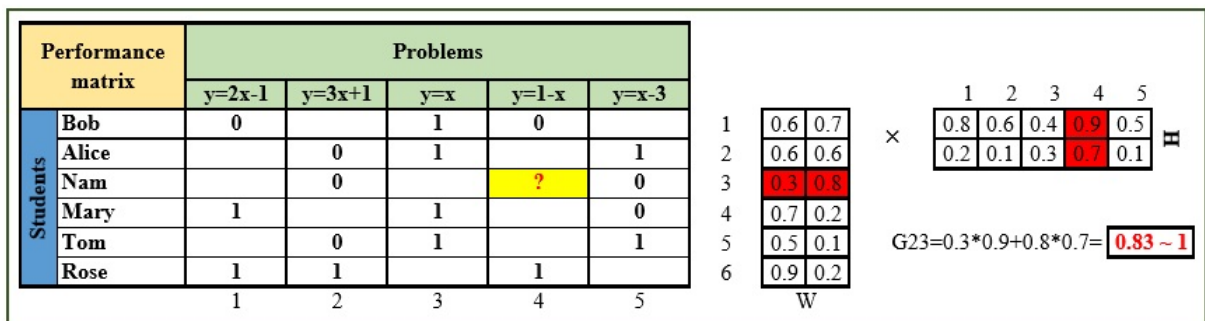
Tìm hiểu và chuẩn bị dữ liệu (Data Understanding Preparation)

Trong nghiên cứu khai thác dữ liệu có nhiều tập dữ liệu sẵn có thường được dùng để thực nghiệm cũng như dữ liệu đánh giá trong các giải thưởng lớn về khai thác dữ liệu giáo dục, Tập dữ liệu ASSISTments được công bố dành cho cộng đồng nghiên cứu về khai thác dữ liệu giáo dục [70]. Tập dữ liệu được phát hành bởi ASSISTments Platform, là hệ thống trợ giảng thông minh trên nền web, giúp giáo viên đánh giá quá trình học môn toán của sinh viên. Hệ thống cho phép giáo viên viết những bài tập riêng, mỗi bài tập bao gồm câu hỏi, video hướng dẫn, đáp án v.v... Cấu trúc dạng phân cấp được biểu diễn như sau: $Assignment \supseteq Assisment \supseteq Problem$

Từ cấu trúc này, chúng ta dễ dàng nhận ra rằng mỗi công việc có một hoặc nhiều chỉ dẫn, mỗi chỉ dẫn có nhiều vấn đề. Nhiệm vụ của chúng ta là dự đoán đúng năng lực (kết quả) của sinh viên khi lần đầu sinh viên giải quyết vấn đề. Tập dữ liệu ASSISTments sau khi xử lý bao gồm 8519 sinh viên (8519 users), 35978 bài tập (35978 items), 1011079 đánh giá năng lực (1011079 ratings).

Mô hình hóa (Modelling)

Hệ thống trợ giảng thông minh ASSITSment được chuyển sang mô hình phân rã ma trận của hệ thống gợi ý được trình bày như (Hình 2.6). Giả sử có 6 sinh viên và 5 bài tập được trình bày trên một ma trận X . Ví dụ, bài tập của sinh viên là tính giá trị y từ giá trị x theo hàm $y = 2x - 1$. Năng lực của sinh viên được đánh giá với thang điểm là 1 (đúng) hoặc 0 (sai). Vấn đề là dự đoán những bài tập mà sinh viên chưa làm (những ô trống trong ma trận X)



Hình 2.6: Phân rã mô hình dự đoán trên tập dữ liệu ASSISTments

Đánh giá mô hình (Evaluation)

Để đánh giá hiệu quả của giải thuật, chúng tôi sử dụng nghi thức kiểm tra hold-out:

lấy ngẫu nhiên 2/3 tập dữ liệu để học và 1/3 còn lại để kiểm tra.

Tuy nhiên do bài toán dự đoán kết quả học tập của sinh viên thuộc dạng dự đoán từ đánh giá tường minh (rating prediction), nên có hai cách đánh giá phù hợp nhất là Root Mean Squared Error (RMSE) và Mean Absolute Error (MAE).

Triển khai mô hình dự đoán (Deployment)

Để dự đoán thành tích học tập của sinh viên, kỹ thuật phân rã ma trận (Matrix factorization - MF) được đề xuất. Trong MF, sinh viên được coi như người dùng, mỗi môn học là một mục tin, điểm của sinh viên được coi là đánh giá và dự đoán sẽ là số điểm (trên thang điểm 10) cho từng môn. Thay vì sử dụng độ chính xác hay AUC để đánh giá, hệ số tương quan và độ sai lệch trung bình tuyệt đối được lựa chọn làm tiêu chí đánh giá. Điểm đặc biệt trong nghiên cứu này là chỉ dự đoán điểm cho mỗi môn học dựa trên mối tương tác giữa sinh viên và môn, không cần thông tin nhân khẩu học. Tuy nhiên, phương pháp này cũng có khả năng được áp dụng để dự đoán điểm theo học kỳ như các nghiên cứu trước.

2.3.2 Dự đoán quy luật chung (không cá nhân hóa)

Để giúp sinh viên tránh nguy cơ bị cảnh cáo về học vụ và rủi ro bị đình chỉ học, thông tin nhân khẩu học từ hồ sơ nhập học của họ, như độ tuổi, giới tính, chuyên ngành, trình độ tiếng nước ngoài và điểm trung bình của học kỳ trước, được sử dụng để dự báo điểm cho học kỳ tiếp theo. Để thực hiện dự báo này, các kỹ thuật khai thác dữ liệu thích hợp cần được áp dụng. Hai giải thuật thường được lựa chọn là cây quyết định và mạng Bayes. Tiếp theo, hiệu suất của cả hai giải thuật sẽ được so sánh. Quá trình khai thác dữ liệu cho hệ thống dự đoán khả năng học tập của sinh viên gồm các bước: tìm hiểu vấn đề, tìm hiểu và chuẩn bị dữ liệu, mô hình hóa, đánh giá mô hình và cuối cùng là triển khai.

Tìm hiểu vấn đề (Business understanding)

Mục tiêu chính là dự đoán kết quả học tập của sinh viên trong một học kỳ cụ thể bằng cách sử dụng thông tin nhân khẩu học (bao gồm độ tuổi, giới tính, trình độ ngoại ngữ,...) và kết quả học tập của học kỳ trước. Điều này giúp cho sinh viên tự đánh giá năng lực

của mình và lên kế hoạch học tập phù hợp, đồng thời cũng giúp cho các giáo viên cố vấn học tập có thể sớm nhận ra các sinh viên có nguy cơ đạt kết quả thấp.

Tìm hiểu và chuẩn bị dữ liệu (Data Understanding Preparation)

Nhằm xây dựng mô hình dự đoán chính xác, nhóm nghiên cứu [71] đã tiến hành tìm hiểu và thu thập dữ liệu từ hệ thống thực tế của Trường Đại học Cần Thơ. Sau khi thu thập, dữ liệu đã được tiền xử lý để loại bỏ các giá trị dư thừa và thiếu, với tổng số 20492 mẫu tin tương ứng với 20492 sinh viên. Tuy nhiên, tập dữ liệu này có rất nhiều thuộc tính, không biết thuộc tính nào sẽ có ảnh hưởng lớn đến kết quả dự đoán. Để giải quyết vấn đề này, nhóm đã sử dụng phương pháp lựa chọn thuộc tính trong công cụ WEKA để đánh giá độ lợi thông tin của từng thuộc tính bằng phương pháp "Information Gain Attribute Evaluation". Từ đó, một ngưỡng được lựa chọn để loại bỏ các thuộc tính không quan trọng, chỉ giữ lại 14 thuộc tính quan trọng để sử dụng trong quá trình dự đoán kết quả học tập của sinh viên. Ví dụ, để dự đoán kết quả học kỳ 5 cho sinh viên s , nhóm sẽ sử dụng các thuộc tính quan trọng đã được xác định trước đó để loại bỏ các thuộc tính ít quan trọng và tạo ra một tập dữ liệu thu gọn để sử dụng cho quá trình dự đoán.

Mô hình hóa (Modelling)

Sau khi tiền xử lý dữ liệu, giai đoạn mô hình hóa bao gồm các kỹ thuật sử dụng, biến đầu vào và biến dự đoán (target attribute) cho việc dự đoán. Kết quả dự đoán có thể chia thành hai loại: phân lớp (dùng cho dự đoán các nhóm điểm chữ như A, B+ hay Xuất sắc, Giỏi,...) và hồi quy (dùng cho dự đoán điểm số như 3.25, 3.15,...).

Đánh giá mô hình (Evaluation)

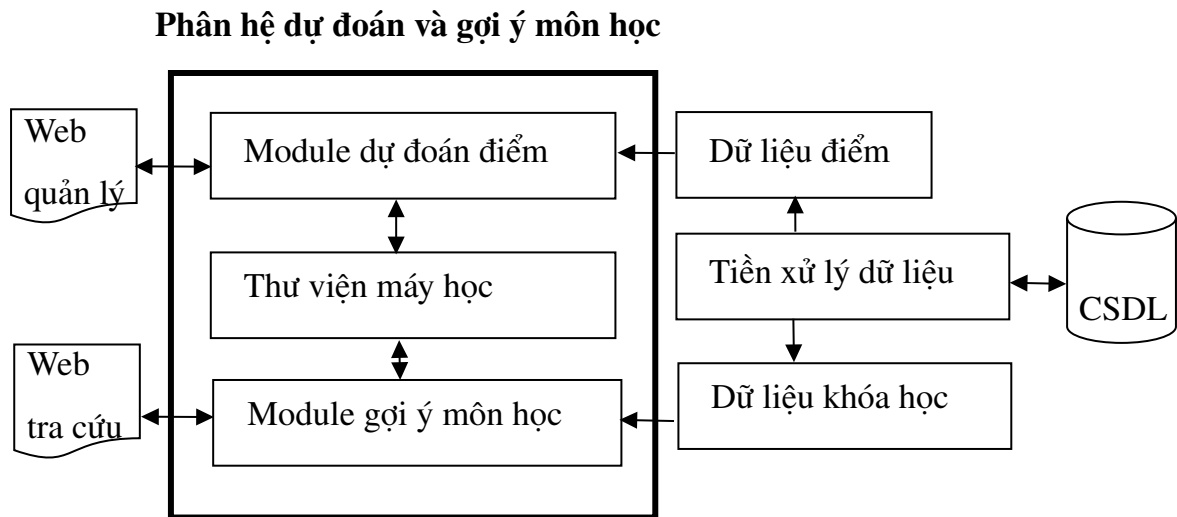
Sau khi xây dựng mô hình, việc đánh giá độ chính xác của các mô hình sẽ được thực hiện bằng cách thực hiện các thao tác tinh chỉnh các thuộc tính và thay đổi các tham số để tìm ra mô hình tốt nhất. Kết quả sẽ được so sánh giữa hai phương pháp: cây quyết định và mạng Bayes. Để đảm bảo tính chính xác của kết quả, phương pháp kiểm tra chéo 10 lần (10-folds cross validation) sẽ được sử dụng.

Để đánh giá mô hình trong bài toán dự đoán giá trị kiểu điểm số (hồi quy) sử dụng độ lỗi (Error) làm phép đo thay vì sử dụng độ chính xác hay AUC. Có nhiều phương pháp

để đo lỗi và đánh giá mô hình, nhưng hai phương pháp phổ biến nhất là hệ số tương quan (Correlation Coefficient) và độ sai lệch trung bình tuyệt đối (Mean Absolute Error).

Triển khai (Deployment)

Hình 2.7 mô tả cấu trúc tổng quát của hệ thống hỗ trợ dự đoán kết quả học tập trên nền web. Công nghệ được sử dụng là JSP hoặc Servlet với sự hỗ trợ của thư viện Weka.



Hình 2.7: Mô hình tổng quát cho hệ thống dự đoán

2.3.3 Các giải pháp gợi ý của cố vấn học tập

Tư vấn hoặc cố vấn học tập là một phần quan trọng để đảm bảo rằng hệ thống đưa ra các gợi ý phù hợp nhất cho người dùng. Trong trường hợp tư vấn học tập cho sinh viên, mục tiêu của hệ thống gợi ý là giúp sinh viên nâng cao kiến thức và kỹ năng trong môn học, mục tiêu của hệ thống gợi ý là giúp giảng viên cải thiện phương pháp giảng dạy và nâng cao chất lượng đào tạo.

Trong mô hình gợi ý thường được thực hiện bằng cách sử dụng các thuật toán máy học và trí tuệ nhân tạo để phân tích dữ liệu về lịch sử học tập của sinh viên, hoặc dữ liệu về giảng dạy của giảng viên. Dữ liệu này có thể bao gồm các thông tin về độ khó của môn học, sở thích và lựa chọn của sinh viên, hoặc phản hồi của sinh viên về phương pháp giảng dạy của giảng viên. Các phương pháp cố vấn học tập trong mô hình gợi ý có thể bao gồm:

Gợi ý nội dung: Hệ thống gợi ý nội dung sẽ dựa trên lịch sử tìm kiếm, đánh giá và sở thích của người dùng để đưa ra các nội dung phù hợp với từng người dùng, ví dụ như

phim, sách, bài viết, tin tức,...

Gợi ý bạn bè: Hệ thống gợi ý bạn bè sẽ dựa trên các thông tin mạng xã hội để đưa ra các đề xuất kết bạn phù hợp với sở thích và khu vực của người dùng.

Gợi ý tài liệu học tập: Hệ thống gợi ý có thể đề xuất cho sinh viên các tài liệu học tập phù hợp với môn học, trình độ và sở thích của họ.

Gợi ý hoạt động ngoại khóa: Hệ thống gợi ý có thể đưa ra các hoạt động ngoại khóa phù hợp với sở thích và sở trường của sinh viên.

Gợi ý các môn học tương tự: Hệ thống gợi ý có thể đề xuất cho sinh viên các môn học tương tự hoặc liên quan đến môn học hiện tại của họ.

Gợi ý các trải nghiệm học tập khác: Hệ thống gợi ý có thể đề xuất cho sinh viên các trải nghiệm học tập khác như thực tập, tham gia dự án, hoặc học tập ở nước ngoài.

2.4 Tổng kết chương

Chương này đã cung cấp một tổng quan về cơ sở lý thuyết và các hướng nghiên cứu liên quan đến việc sử dụng giải pháp gợi ý trong cố vấn học tập. Nội dung bao gồm cấu trúc của hệ thống trợ giảng thông minh và cách áp dụng nó vào bài toán dự đoán năng lực học tập của sinh viên. Từ đó, chương trình đưa ra các gợi ý phù hợp với từng sinh viên cụ thể để hỗ trợ việc lập kế hoạch học tập, đặc biệt là trong việc lựa chọn môn học tự chọn. Các nghiên cứu trên lĩnh vực này đã sử dụng nhiều phương pháp khác nhau để đưa ra các gợi ý phù hợp, bao gồm cả phương pháp dựa trên năng lực học tập và phương pháp dựa trên lịch sử học tập của sinh viên. Đồng thời, các nghiên cứu này cũng chú trọng đến việc đánh giá hiệu quả của các giải pháp gợi ý và cách cải thiện chúng để tăng tính chính xác.

Một hệ thống trợ giảng thông minh có kiến trúc nền tảng gồm bốn mô hình cơ bản: mô hình người học, mô hình giảng dạy, mô hình kiến thức và giao diện người dùng. Trong số các mô hình này, mô hình người học được coi là cốt lõi và đạt được nhiều thành tựu trong lĩnh vực trí tuệ nhân tạo và khai thác dữ liệu.

Hệ thống gợi ý đang phát triển mạnh mẽ trên nhiều lĩnh vực, bao gồm giải trí, du lịch, thương mại, nông nghiệp và giáo dục. Trong đó, hệ thống gợi ý đang nhận được sự quan tâm đặc biệt từ các nhà nghiên cứu giáo dục và ngày càng được áp dụng rộng rãi trong

giáo dục. Chương này cung cấp một tổng quan về hệ thống gợi ý, bao gồm quy trình xây dựng hệ thống và những hướng nghiên cứu tiềm năng.

Bài toán dự đoán năng lực học tập của sinh viên có thể được phân loại thành hai nhóm chính, tùy thuộc vào cách tiếp cận. Nhóm một là dự đoán không cá nhân hóa, trong đó kết quả học tập của từng sinh viên được xem là một phần của một tập dữ liệu lớn hơn. Hướng tiếp cận này đưa ra những dự đoán chung chung và quy luật, dựa trên phân tích các yếu tố chung như độ tuổi, giới tính, ngành học, điểm trung bình, v.v. của nhóm sinh viên. Nhóm hai là dự đoán cá nhân hóa, trong đó mỗi sinh viên được xem là một thực thể độc lập, có các đặc trưng riêng như hồ sơ, quá trình học tập, sở thích, v.v. Hướng tiếp cận này đưa ra dự đoán cụ thể cho từng sinh viên, dựa trên việc áp dụng các thuật toán học máy và phân tích các đặc trưng độc lập của mỗi sinh viên. Vì vậy, dự đoán cá nhân hóa có thể cho kết quả chính xác hơn và giúp tăng cường hiệu quả giảng dạy cũng như học tập của từng cá nhân.

Dựa trên cơ sở lý thuyết và các nghiên cứu liên quan, luận án đã tổng quan bài toán giải pháp gợi ý trong việc cố vấn học tập cho sinh viên. Tuy nhiên, để đạt được độ chính xác cao và đáp ứng nhu cầu thực tế, nghiên cứu đã đưa ra một số phương pháp cải tiến hệ thống gợi ý như sau: (1) Áp dụng dự đoán kết quả học tập sinh viên theo hướng hệ thống gợi ý và đồng thời giải quyết các vấn đề không đồng nhất trong đánh giá, được trình bày ở chương 3; (2) Áp dụng kiến trúc học sâu vào mô hình dự đoán để cải thiện hiệu suất dự đoán đã trình bày trong chương 4; (3) Đề xuất phương pháp tích hợp mối quan hệ giữa người dùng cũng như mối liên quan các môn học trong việc dự đoán kết quả học tập của sinh viên, chi tiết được trình bày trong chương 5.

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

Chương này trình bày một số kết quả nghiên cứu trong việc áp dụng hệ thống gợi ý để dự đoán kết quả học tập cá nhân hóa cho từng sinh viên. Điểm độc đáo của nghiên cứu này là việc tập trung vào dự đoán kết quả riêng lẻ của mỗi sinh viên thay vì dựa vào dự đoán quy luật chung như các nghiên cứu trước. Thông qua thực nghiệm, kết quả cho thấy phương pháp mới này mang lại độ chính xác cao và rất tiềm năng để được tích hợp trong hệ thống gợi ý môn học, hỗ trợ quá trình tư vấn học tập của cố vấn học tập. Bên cạnh đó, chương này cũng giới thiệu việc tích hợp giá trị độ lệch vào mô hình, nhằm giải quyết vấn đề đánh giá không đồng đều và nâng cao độ chính xác của dự đoán (thách thức đã được nêu trong mục 1.3). Nội dung chương này được tổng hợp từ công trình nghiên cứu đã công bố [CT1].

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

3.1.1 Phát biểu bài toán hệ thống gợi ý trong ngữ cảnh giáo dục

Bài toán gợi ý này, còn được gọi là bài toán gợi ý hai chiều truyền thống [47], có thể được hình thức hóa bởi:

Cho tập hợp hữu hạn gồm N sinh viên $S = \{s_1, s_2, \dots, s_N\}$ và M môn học $C = \{c_1, c_2, \dots, c_M\}$

Mỗi sinh viên $s \in S$ được biểu diễn thông qua $|P|$ đặc trưng $P = \{p_1, p_2, \dots, p_{|P|}\}$. Các đặc trưng $p \in P$ thông thường là thông tin cá nhân của mỗi sinh viên (Student Profile). Ví dụ $s \in S$ là một sinh viên thì các đặc trưng biểu diễn cho sinh viên s có thể là $P = \{\text{giới tính, tuổi, lớp, điểm trung bình,} \dots\}$.

Mỗi môn học $c \in C$ cũng có thể là bài tập, bài kiểm tra, bài thi hoặc bất kỳ dạng tài nguyên nào mà sinh viên cần đến. Mỗi môn học $c \in C$ được biểu diễn thông qua $|A|$ đặc

trưng nội dung $A = \{a_1, a_2, \dots, a_{|A|}\}$. Các đặc trưng $a_A \in A$ nhận được từ các phương pháp trích chọn đặc trưng trong lĩnh vực truy vấn thông tin. Ví dụ $c \in C$ là một môn học thì các đặc trưng nội dung biểu diễn trong môn học c có thể là $A = \{\text{ngành, nhóm môn học, số tín chỉ, nội dung, } \dots\}$.

Mối quan hệ giữa tập sinh viên S và tập môn học C được biểu diễn thông qua ma trận đánh giá năng lực sinh viên $G = [g_{sc}]$ với $s = 1, 2, \dots, N; c = 1, 2, \dots, M$. Hình 3.1 thể hiện N sinh viên đã học M môn học, điểm số là giá trị của các ô trong ma trận đánh giá năng lực, và một sinh viên a đã học $M - 1$ môn học và cần dự đoán điểm của 1 môn học mà sinh viên này chưa được học.

		Môn học					
		1	2		c		M
Sinh viên	1	3	0	2	1	4	4
	2	4	0	3	2	1	4
							
	s	1	4	0	4	3	2
							
	N	4	1	1	0	2	3
							
	a	0	4	0	1	0	1
		0	3	0	3	4	0
		0	3	?	3	4	0

Hình 3.1: Ví dụ ma trận đánh giá năng lực sinh viên

Giá trị thể hiện năng lực của sinh viên $s \in S$ cho một số môn học $c \in C$. Thông thường giá trị g_{sc} nhận một giá trị thuộc miền $G = \{0, 1, 2, 3, 4\}$ được thu thập trực tiếp bằng điểm số trung bình môn học của sinh viên hoặc thu thập gián tiếp thông qua hoạt động học tập của sinh viên. Những giá trị g_{sc} là năng lực sinh viên $s \in S$ học môn học $c \in C$, những ô điền ký tự "?" được hiểu là sinh viên chưa học môn học đó, nó là giá trị cần hệ thống gợi ý đưa ra dự đoán điểm số. Tiếp đến, ta ký hiệu $C_s \subseteq C$ là tập các môn học $c \in C$ được học và có điểm số bởi sinh viên $s \in S$ và $s_a \in S$ được gọi là sinh viên hiện thời hoặc sinh viên cần được gợi ý. Khi đó, tồn tại hai dạng bài toán điển hình của hệ thống gợi ý là:

- (1) Dự đoán điểm số của sinh viên với các môn học mà sinh viên chưa học.
- (2) Gợi ý danh sách ngắn các môn học phù hợp với sinh viên hiện thời. Cụ thể đối với sinh viên a , hệ thống gợi ý sẽ chọn ra K môn học mới $c \in (C \setminus C_a)$ phù hợp với sinh

viên a nhất để gợi ý cho họ.

3.1.2 Quy trình xây dựng hệ thống gợi ý

Một hệ thống gợi ý thực hiện những bước xử lý chính sau [24]: (1) Thu thập thông tin; (2) Xây dựng mô hình; (3) Dự đoán đánh giá năng lực và đưa ra gợi ý.

Giai đoạn 1: Thu thập thông tin

Hệ thống gợi ý thường thu thập ba loại thông tin chính gồm có:

- Thông tin cá nhân của sinh viên (Student): Được biểu diễn qua các thuộc tính và cho phép hệ thống xây dựng hồ sơ sinh viên để lưu trữ các dấu vết nội dung môn học đã từng được sinh viên học.
- Thông tin về môn học (Course): Được biểu diễn qua các đặc trưng và cho phép hệ thống xây dựng đặc trưng nội dung môn học để lưu trữ các dấu vết các thuộc tính sinh viên đã từng học môn học.
- Năng lực của sinh viên với môn học (Mark/Score): Biểu diễn qua các giá trị năng lực của sinh viên, bao gồm cả năng lực tường minh (điểm số) và năng lực ngầm định (đánh giá qua các hoạt động học tập, có tham gia, thời gian tham gia). Các giá trị đánh giá này cung cấp thông tin quan trọng để hệ thống đưa ra các gợi ý phù hợp với sở trường và năng lực của sinh viên.

Giai đoạn 2: Xây dựng mô hình

Giai đoạn xây dựng mô hình gợi ý bao gồm quá trình thu thập thông tin và các hướng tiếp cận khác nhau để so sánh và đánh giá mối liên hệ giữa các thông tin thu thập được. Các hướng tiếp cận phổ biến bao gồm sử dụng heuristics dựa trên kinh nghiệm, học máy và lý thuyết xấp xỉ,... [72] Mỗi hướng tiếp cận khai thác thông tin đầu vào theo cách riêng để hình thành các phương pháp gợi ý khác nhau.

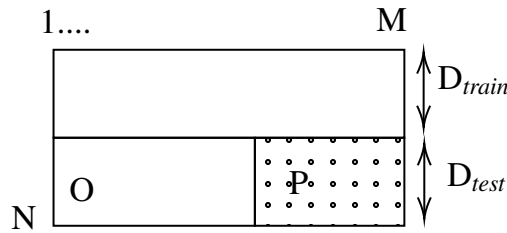
Giai đoạn 3: Dự đoán đánh giá / Đưa ra gợi ý

Sau khi hoàn thành giai đoạn xây dựng mô hình, dữ liệu đầu ra sẽ được sử dụng để dự đoán đánh giá năng lực của sinh viên với các môn học chưa được học trước đó, từ đó đề xuất môn học mới phù hợp nhất với sinh viên hiện tại để cung cấp gợi ý cho họ.

3.1.3 Đánh giá hệ thống gợi ý

Để đảm bảo hiệu quả của hệ thống gợi ý, ta cần sử dụng một mô hình thống kê hoặc mô hình học máy phù hợp nhất. Tuy nhiên, trong quá trình lựa chọn, ta cần đánh giá độ chính xác của từng mô hình để tìm ra mô hình tốt nhất. Vấn đề đặt ra là làm thế nào để đánh giá chính xác và đưa ra quyết định lựa chọn, trong bối cảnh có rất nhiều mô hình khác nhau hiện nay.

Để đánh giá độ chính xác của hệ thống gợi ý, chúng ta cần phải chia ma trận đánh giá G thành hai phần D_{train} và D_{test} sao cho $D_{train} \cup D_{test} = G$ và $D_{train} \cap D_{test} = \emptyset$. Dữ liệu huấn luyện D_{train} được sử dụng để xây dựng mô hình gợi ý, trong khi dữ liệu kiểm tra D_{test} được sử dụng để kiểm tra độ chính xác của mô hình. Việc này giúp đánh giá khả năng áp dụng của các thuật toán lọc trong hệ thống gợi ý (Hình 3.2).



Hình 3.2: Phương pháp phân chia tập dữ liệu phục vụ cho đánh giá hệ thống gợi ý

Dưới đây là các phương pháp chia tập đánh giá năng lực G thành hai tập con: D_{train} và D_{test} theo [29]:

- **Phân chia (Splitting):** Ngẫu nhiên lựa chọn một số hàng từ ma trận đánh giá G và gán cho tập huấn luyện D_{train} . Các hàng còn lại sẽ thuộc về tập thực nghiệm D_{test} .

- **Lấy mẫu (Bootstrap):** Dựa trên một tập dữ liệu ban đầu, ta sử dụng kỹ thuật lấy mẫu có hoàn lại để tạo ra các mẫu mới cho tập huấn luyện D_{train} . Mẫu không được lấy sẽ thuộc về tập thực nghiệm D_{test} .

- **Kiểm thử chéo (k-fold cross validation):** Tập đánh giá G sẽ được phân thành k tập con có kích thước tương đương. Thuật toán gợi ý sẽ được kiểm tra k lần, trong đó mỗi lần sẽ chọn 1 trong k tập con làm tập kiểm tra D_{test} và phần còn lại làm tập huấn luyện D_{train} . Kết quả từ k lần kiểm tra sẽ được lấy trung bình để đánh giá tổng thể.

Sau khi chia tập đánh giá G thành hai phần D_{train} và D_{test} , để đánh giá hệ thống gợi ý hiện thời ta tiến hành như sau: Với mỗi sinh viên $s \in D_{test}$, các đánh giá $g_{s,c} \neq \emptyset$ được

chia thành hai phần O và P . Tập O được coi là tập đã đánh giá (sinh viên đã học và có điểm), trong khi đó tập P là tập đánh giá cần dự đoán từ dữ liệu huấn luyện D_{train} và O .

Giả sử phương pháp lọc đưa ra dự đoán cho sinh viên trong tập P là $\hat{P}$. Khi đó kết quả đánh giá cho hệ thống gợi ý hiện thời sẽ dựa trên sự so sánh các đánh giá từ hai tập P và $\hat{P}$.

Hệ thống gợi ý trong giáo dục có hai nhiệm vụ chính: dự đoán giá trị đánh giá (điểm số dự đoán) và đề xuất một danh sách (top-n) các môn học phù hợp cho sinh viên hiện tại. Có hai nhóm độ đo để đánh giá hệ thống gợi ý, tương ứng với hai nhiệm vụ trên: (1) Nhóm đầu tiên đánh giá độ chính xác của dự đoán giá trị đánh giá; (2) Nhóm thứ hai đánh giá độ chính xác của danh sách các môn học được đề xuất. Cả hai nhóm đều sử dụng thông tin từ ma trận đánh giá (P) và ma trận dự đoán ($\hat{P}$) để tính toán các độ đo. Do bài toán dự đoán kết quả học tập của sinh viên thuộc nhóm đánh giá độ chính xác của dự đoán (nhóm 1) và luận án sẽ trình bày chi tiết về độ đo này ở phần tiếp theo.

3.1.4 Độ đo đánh giá độ chính xác của dự đoán kết quả học tập

Khi đánh giá một hệ thống gợi ý, việc lựa chọn độ đo phù hợp để xác định độ chính xác của dự đoán là cần thiết. Đối với bài toán dự đoán kết quả học tập, chúng ta thường dựa vào sự chênh lệch giữa giá trị điểm dự đoán và giá trị điểm thực tế. Do dự đoán kết quả học tập là dự đoán một giá trị số thực trong khoảng giá trị liên tục nên việc sử dụng các độ đo độ lỗi của dự đoán phù hợp hơn đánh giá dự đoán có giá trị rời rạc.

Độ đo trung bình giá trị tuyệt đối lỗi (Mean Absolute Error - MAE) là một phép đo rất phổ biến, cung cấp giá trị trung bình của sai số tuyệt đối giữa giá trị dự đoán và thực tế.

Sai số dự đoán MAE_s với mỗi sinh viên s thuộc tập dữ liệu kiểm tra D_{test} được tính bằng trung bình cộng của sai số tuyệt đối giữa hai giá trị được dự đoán và giá trị thực của sinh viên s với tất cả các môn học thuộc tập C theo công thức (3.1).

$$MAE_s = \frac{1}{|C|} \sum_{c \in C} |\hat{g}_{sc} - g_{sc}| \quad (3.1)$$

Sai số dự đoán trên toàn tập dữ liệu kiểm tra được tính bằng trung bình cộng sai số

dự đoán cho mỗi sinh viên thuộc D_{test} được tính theo công thức (3.2).

$$MAE = \frac{\sum_{s \in D_{test}} MAE_s}{|D_{test}|} \quad (3.2)$$

Độ đo lỗi bình phương trung bình gốc (Root Mean Square Error - RMSE) là phép đo được sử dụng rộng rãi, đánh giá bằng cách lấy căn bậc hai của trung bình các bình phương sai số giữa giá trị dự đoán và thực tế. RMSE cho từng sinh viên được tính theo công thức (3.3) và RMSE cho cả tập dữ liệu kiểm tra theo công thức (3.4).

$$RMSE_s = \sqrt{\frac{1}{|C|} \sum_{c \in C} (\hat{g}_{sc} - g_{sc})^2} \quad (3.3)$$

$$RMSE = \frac{\sum_{s \in D_{test}} RMSE_s}{|D_{test}|} \quad (3.4)$$

Tùy theo mục đích của việc đánh giá, nếu muốn xem xét đến sự ảnh hưởng của những sai số lớn thì RMSE sẽ là lựa chọn tốt hơn do nó phạt nặng hơn cho những sai số lớn. Trong trường hợp muốn xem xét sự chênh lệch trung bình, không quan tâm đến sự phân bố của sai số, MAE sẽ là phép đo thích hợp. Một hệ thống gợi ý hiệu quả sẽ có giá trị MAE và RMSE thấp, chỉ ra rằng dự đoán của nó gần với giá trị thực tế.

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

Phương pháp lọc cộng tác dựa trên người dùng [5] đã đem lại những thành tựu đáng kể trong việc dự đoán đánh giá và ứng dụng trong nhiều lĩnh vực. Ý tưởng căn bản của phương pháp này tương tự như việc bạn nhận xét và tham khảo ý kiến từ những người có sở thích và hành vi tương tự như bạn khi quyết định về một vấn đề nào đó. Trong trường hợp của hệ thống dự đoán năng lực học tập, chúng ta tập trung vào việc sử dụng toàn bộ ma trận đánh giá, thể hiện sự tương tác giữa các sinh viên và các môn học, để xây dựng một mô hình phân tích và tiên đoán năng lực học tập.

Ví dụ, trong phương pháp lọc cộng tác dựa trên người dùng, chúng ta xem xét những sinh viên có các sở thích học tương tự nhau. Điều này dựa trên giả định rằng những

người có sở thích tương tự sẽ có cách tiếp cận và hiểu biết tương đương về những môn học tương tự. Khi kết hợp đánh giá từ những sinh viên tương tự, chúng ta có thể tạo ra dự đoán về năng lực học tập của một sinh viên cụ thể trên những môn học mà họ chưa được đánh giá.

Mô hình lọc cộng tác này có thể áp dụng không chỉ trong lĩnh vực dự đoán năng lực học tập mà còn trong nhiều bài toán khác. Tương tự như việc bạn cân nhắc ý kiến của những người có kinh nghiệm hoặc sở thích tương tự khi bạn đứng trước một quyết định quan trọng, việc tìm ra những người có đặc điểm tương đồng trong một hệ thống và sử dụng thông tin từ họ để dự đoán, đề xuất hoặc tư vấn có thể mang lại những kết quả ấn tượng.

Trong bối cảnh giáo dục, việc áp dụng phương pháp lọc cộng tác dựa trên người dùng và môn học có thể giúp chúng ta hiểu rõ hơn về năng lực học tập của từng sinh viên và từ đó thiết kế các môn học, kế hoạch học tập, hoặc các phương pháp giảng dạy phù hợp với từng cá nhân. Điều này không chỉ giúp tối ưu hóa quá trình học tập mà còn đóng góp vào việc phát triển cá nhân toàn diện cho từng học sinh.

Để tiếp cận bài toán dự đoán năng lực học tập của sinh viên thông qua phương pháp lọc cộng tác, dựa trên những lý thuyết đã trình bày, ta đặt ra giả định rằng những sinh viên có điểm số tương tự sẽ có khả năng học tương tự trên các môn học có đặc điểm tương đương. Điều này thúc đẩy việc áp dụng phương pháp lọc cộng tác dựa trên người dùng và môn học như một lựa chọn đáng xem để giải quyết bài toán dự đoán năng lực học tập của sinh viên. Phương pháp này, trong ngữ cảnh giáo dục, thường được gọi là Student k -NNs, với ý tưởng chính là sử dụng k láng giềng gần nhất. Cụ thể, phương pháp này được thực hiện qua bốn bước chính:

Bước 1. Tính toán mức độ tương tự của tất cả sinh viên trong hệ thống với sinh viên hiện thời (u_a).

Mức độ tương tự giữa sinh viên $s' \in S$ với sinh viên hiện thời s , kí hiệu là $\text{sim}(s, s')$, được xem xét dựa vào tập các môn học mà cả hai sinh viên này đều đã học. Có nhiều độ đo khác nhau tính toán mức độ tương tự như: Độ đo khoảng cách (Euclid, Minkowski); Độ đo tương tự (Cosine, Entropy); Độ đo tương quan (Pearson, Spearman, Kendal). Trong đó hai độ đo phổ biến thường hay được sử dụng là độ tương quan Pearson và giá trị Cosine giữa hai véc-tơ.

Độ tương tự Cosine giữa hai sinh viên bởi s và s' : là giá trị Cosine giữa hai véc-tơ bởi s và s' theo công thức (3.5). Trong đó, bởi s và s' được biểu diễn bằng véc-tơ m chiều (và $n = |C_{ss'}|$ là số lượng các môn học cả hai sinh viên đã học).

$$\text{sim}_{\text{cosine}}(s, s') = \frac{\sum_{c \in C_{ss'}} g_{s'c} \cdot g_{sc}}{\sqrt{\sum_{c \in C_{ss'}} g_{sc}^2 \sum_{c \in C_{ss'}} g_{s'c}^2}} \quad (3.5)$$

Tương tự như vậy, độ tương quan Pearson giữa hai sinh viên được tính theo công thức (3.6) được biểu diễn như sau:

$$\text{sim}_{\text{pearson}}(s, s') = \frac{\sum_{c \in C_{ss'}} (g_{sc} - \bar{g}_s) (g_{s'c} - \bar{g}_{s'})}{\sqrt{\sum_{c \in C_{ss'}} (g_{sc} - \bar{g}_s)^2 \sum_{c \in C_{ss'}} (g_{s'c} - \bar{g}_{s'})^2}} \quad (3.6)$$

Trong đó:

- $C_{ss'} = \{c \mid g_{sc} \neq \emptyset \cap g_{s'c} \neq \emptyset\}$ là tập tất cả các môn học cùng được học bởi s và s'
- $\bar{g}_s, \bar{g}_{s'}$ là trung bình cộng các điểm khác $\emptyset$ của s và s'

Bước 2. Xác định tập sinh viên láng giềng với s bằng việc chọn K_1 sinh viên có mức độ tương tự cao nhất với s .

Bước 3. Sinh ra dự đoán điểm học tập của s với môn học chưa học bằng việc kết hợp các điểm số của các sinh viên trong tập láng giềng.

Gọi $\hat{S}$ là tập K_1 sinh viên tương tự nhất đối với s (Tập láng giềng với s). Khi đó mức độ phù hợp của sinh viên s với môn học mới c được xác định như một hàm các đánh giá của tập láng giềng theo một số phương pháp phổ biến được tính theo công thức (3.7) (3.8) và (3.9):

$$\hat{g}_{sc} = \frac{1}{K_1} \sum_{s' \in \hat{S}} g_{s'c} \quad (3.7)$$

$$\hat{g}_{sc} = h \sum_{s' \in \hat{S}} \text{sim}(s, s') \times g_{s'c} \quad (3.8)$$

$$\hat{g}_{sc} = \bar{g}_s + h \sum_{s' \in \hat{S}} \text{sim}(s, s') \times (g_{s'c} - \bar{g}_{s'}) \quad (3.9)$$

Trong đó: h được gọi là nhân tố chuẩn hóa theo công thức (3.10), $\bar{g}_s$ là trung bình các đánh giá của sinh viên s được xác định theo công thức (3.11).

$$h = 1 / \sum_{s' \in \hat{S}} \text{sim}(s, s') \quad (3.10)$$

$$\bar{g}_s = \frac{1}{|C_c|} \sum_{c \in C_c} g_{sc} \quad (3.11)$$

Trong đó: $C_c = \{c \in C_c \mid g_{sc} \neq 0\}$

Bước 4. Gợi ý K_2 môn học mới sẽ điểm cao nhất cho s .

Luận bàn về trường hợp sử dụng của hai độ đo tương tự.

Độ đo Cosine và Pearson là hai phương pháp tương tự được sử dụng rộng rãi trong hệ thống gợi ý. Khi áp dụng vào bài toán dự đoán kết quả học tập của sinh viên:

- Độ đo Cosine chủ yếu tính toán sự tương tự giữa hai vector dựa trên góc giữa chúng. Nó phù hợp với dữ liệu thưa và chủ yếu xem xét mức độ giống nhau mà không quan tâm đến sự sai lệch trong cách đánh giá.
- Trong khi đó, Độ đo Pearson tính toán hệ số tương quan giữa hai vector, giúp xem xét sự sai lệch trong cách đánh giá. Ví dụ, nếu hai sinh viên có xu hướng đánh giá khác nhau, nhưng vẫn có cùng một xu hướng, Pearson sẽ coi họ là tương tự.

Trong việc dự đoán kết quả học tập, Cosine có thể hữu ích khi chỉ có ít dữ liệu về điểm số, trong khi Pearson giúp xác định sự tương đồng giữa sinh viên dựa trên xu hướng đánh giá của họ. Lựa chọn giữa hai phương pháp phụ thuộc vào bản chất của dữ liệu và mục tiêu của hệ thống.

Độ phức tạp của giải thuật Student-kNNs

- Xây dựng ma trận tương đồng (similarity matrix): Để tính độ tương đồng giữa các sinh viên, bạn cần tính toán độ tương đồng (similarity) giữa tất cả các cặp sinh viên. Giả sử có s sinh viên và c môn học, độ phức tạp của bước này là khoảng $O(s \times s \times c)$.

- Tìm k sinh viên giống nhau nhất: Với mỗi sinh viên, bạn cần tìm k sinh viên giống nhau nhất. Điều này đòi hỏi sắp xếp và chọn k phần tử, có độ phức tạp cho giải thuật sắp xếp phổ biến là $O(s \times \log(s))$.
- Dự đoán điểm cho môn học cụ thể: Khi đã xác định được k sinh viên giống nhau nhất, ta tính toán điểm dự đoán cho môn học cụ thể. Điều này đòi hỏi $O(k)$ thời gian.

Vì vậy, tổng độ phức tạp thời gian của thuật toán Student kNNs có thể được xấp xỉ là $O(s \times s \times c + s \times \log(s) + k)$, theo nguyên tắc lấy tối đa, độ phức tạp của thuật toán là $O(s \times s \times c)$.

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

Phương pháp dự đoán dựa vào môn học tương tự này có điểm khác biệt so với phương pháp lọc cộng tác dựa trên sinh viên tương tự. Thay vì tính toán mức độ tương tự giữa sinh viên hiện tại s và các sinh viên khác trong hệ thống, phương pháp này tập trung vào việc tính toán mức độ tương tự giữa môn học mục tiêu cần dự đoán và các môn học mà sinh viên s đã học trước đó.

Cách tiếp cận này bắt đầu bằng việc đánh giá mức độ tương tự giữa các môn học. Thay vì so sánh trực tiếp các sinh viên, phương pháp này xem xét các sinh viên đã tham gia cả hai môn học đang được so sánh. Bằng cách làm như vậy, phương pháp tạo ra một tập hợp các môn học "láng giềng" gần với môn học mục tiêu.

Tiếp theo, phương pháp sử dụng tập hợp các môn học láng giềng này để dự đoán điểm số cho sinh viên s trên môn học mục tiêu. Điều này thường bao gồm việc tính toán một trọng số trung bình dựa trên điểm số của sinh viên s trong các môn học láng giềng.

Phương pháp này thực hiện qua bốn bước chính:

Bước 1. Tương tự như việc tính mức độ tương tự giữa hai sinh viên ở phương pháp lọc cộng tác dựa vào sinh viên, việc tính toán mức độ tương tự giữa hai môn học cũng được thực hiện tương tự theo một số độ đo phổ biến đã đề cập là độ tương tự Pearson và giá trị Cosine giữa hai véc-tơ.

Độ tương quan Pearson giữa hai môn học c và c' được tính theo công thức (3.12):

$$\text{sim}_{\text{pearson}}(c, c') = \frac{\sum_{s \in S_{cc'}} (g_{sc} - \bar{g}_c) (g_{sc'} - \bar{g}_{c'})}{\sqrt{\sum_{s \in S_{cc'}} (g_{sc} - \bar{g}_c)^2 \sum_{s \in S_{cc'}} (g_{sc'} - \bar{g}_{c'})^2}} \quad (3.12)$$

Trong đó:

- $S_{cc'} = \{s \mid g_{sc} \neq \emptyset \cap g_{sc'} \neq \emptyset\}$ là tập tất cả các sinh viên đã học cả hai môn học c và môn học c' .
- $\bar{g}_c, \bar{g}_{c'}$ là trung bình cộng các điểm số khác $\emptyset$ cho môn học c và c' .

Độ tương tự Cosine giữa môn học c và c' : là cosine của hai véc-tơ c và c' theo công thức (3.13). Trong đó, hai môn học c và c' được xem xét như véc-tơ c và n chiều ($n = |S_{cc'}|$ là số lượng các sinh viên đã học cả hai môn học c và c').

$$\text{sim}_{\text{cosine}}(c, c') = \cos(c, c') = \frac{\vec{c} \cdot \vec{c'}}{\|\vec{c}\| \|\vec{c'}\|} = \frac{\sum_{s \in S_{cc'}} g_{sc} \cdot g_{sc'}}{\sqrt{\sum_{s \in S_{cc'}} g_{sc}^2} \cdot \sqrt{\sum_{s \in S_{cc'}} g_{sc'}^2}} \quad (3.13)$$

Bước 2. Xác định tập môn học láng giềng với c bằng việc chọn K_1 môn học có mức độ tương tự cao nhất với c

Bước 3. Việc sinh ra dự đoán điểm của sinh viên hiện thời s với các môn học chưa học c được tính toán bằng việc kết hợp các điểm số của s đối với các môn học trong tập láng giềng của c' . Gọi $\hat{C}$ là tập gồm K_1 môn học tương tự nhất đối với c' . Khi đó mức độ phù hợp của s với c' được xác định như một hàm các đánh giá của tập láng giềng. Phương pháp phổ dụng nhất để dự đoán mức độ phù hợp của môn học c' đối với sinh viên s được xác định theo công thức (3.14).

$$\hat{g}_{sc} = \frac{\sum_{c' \in \hat{C}} g_{sc'} \cdot \text{sim}(c, c')}{\sum_{c' \in \hat{C}} |\text{sim}(c, c')|} \quad (3.14)$$

Bước 4. Gọi ý K_2 môn học mới có điểm dự đoán cao nhất cho sinh viên s . Từ kết quả dự đoán các môn học mà sinh viên có năng lực cao sẽ được lựa chọn và đưa vào kế hoạch học tập của mình.

Độ phức tạp của giải thuật Course-kNNs

- Xây dựng ma trận tương đồng (similarity matrix): Để tính độ tương đồng giữa các

môn học, ta cần tính toán độ tương đồng (similarity) giữa tất cả các cặp môn học. Giả sử có s sinh viên và c môn học, độ phức tạp của bước này là khoảng $O(c \times c \times s)$.

- Tìm k môn học giống nhau nhất: Với mỗi môn học, bạn cần tìm k môn học giống nhau nhất. Điều này đòi hỏi sắp xếp và chọn k phần tử, có độ phức tạp cho giải thuật sắp xếp phổ biến là $O(c \times \log(c))$.
- Dự đoán điểm cho môn học cụ thể: Khi đã xác định được k môn học giống nhau nhất, ta tính toán điểm dự đoán cho môn học cụ thể. Điều này đòi hỏi $O(k)$ thời gian.

Vì vậy, tổng độ phức tạp thời gian của thuật toán Course kNNs có thể được xấp xỉ là $O(c \times c \times s + c \times \log(c) + k)$, theo nguyên tắc lấy tối đa, độ phức tạp của thuật toán là $O(c \times c \times s)$.

Hạn chế của 2 phương pháp lọc công tác dựa trên tính toán tương tự

Phương pháp lọc cộng tác dựa trên sinh viên/ môn học tương tự đem lại nhiều lợi ích cho việc dự đoán năng lực học tập của sinh viên, nhưng cũng tồn tại một số hạn chế đáng chú ý, đặc biệt khi gặp phải dữ liệu lớn và dữ liệu thưa.

Dữ liệu lớn đôi khi có thể trở thành một rào cản trong việc thực hiện phương pháp lọc cộng tác dựa trên người dùng tương tự. Khi số lượng sinh viên và môn học tăng lên, quá trình tính toán mức độ tương tự giữa tất cả các cặp sinh viên hoặc môn học trở nên phức tạp và tốn kém về thời gian và tài nguyên. Việc xử lý lượng lớn dữ liệu có thể dẫn đến sự trễ trong việc đưa ra dự đoán và giảm hiệu suất của hệ thống.

Ngoài ra, dữ liệu thưa là một thách thức đối với phương pháp lọc cộng tác này. Trong môi trường học tập, không phải tất cả sinh viên tham gia và đánh giá trên cùng mọi môn học, dẫn đến ma trận đánh giá thưa với nhiều giá trị bị thiếu. Dữ liệu thưa gây khó khăn trong việc xác định các sinh viên hoặc môn học tương tự, và do đó làm giảm độ chính xác của dự đoán.

Để vượt qua các hạn chế này, phương pháp phân rã ma trận áp dụng vào bài toán dự đoán năng lực học tập của sinh viên đã được đề xuất. Phân rã ma trận tập trung vào việc biểu diễn dữ liệu dưới dạng các ma trận nhỏ hơn, mô tả các đặc điểm tiềm ẩn của cả sinh viên và môn học. Việc này giúp giảm tác động của dữ liệu thưa bằng cách tạo ra các dự đoán dựa trên các đặc điểm tiềm ẩn, cải thiện khả năng dự đoán và hiệu suất của hệ thống trong môi trường học tập.

3.4 Phương pháp phân rã ma trận trong dự đoán kết quả học tập sinh viên (PSP-MF)

Phương pháp phân rã ma trận là một kỹ thuật mạnh mẽ và linh hoạt được áp dụng trong việc dự đoán kết quả học tập của sinh viên. Khi áp dụng vào môi trường giáo dục, phương pháp này giúp hiểu và dự đoán năng lực học tập của sinh viên dựa trên thông tin về quá trình học và đánh giá.

Tại gốc của phương pháp là giả định rằng dữ liệu học tập có thể được biểu diễn dưới dạng một ma trận, với hàng tương ứng với sinh viên và cột tương ứng với môn học. Mục tiêu của phương pháp là phân tách ma trận đánh giá ban đầu thành hai ma trận nhỏ hơn, một ma trận chứa thông tin về các yếu tố tiềm ẩn của sinh viên và một ma trận chứa thông tin về các yếu tố tiềm ẩn của môn học.

Cách thức phân tách này dựa trên giả định rằng mỗi sinh viên có những đặc điểm ẩn không thể quan sát trực tiếp, như sự chăm chỉ, khả năng tiếp thu, hay phong cách học tập. Tương tự, mỗi môn học cũng có các yếu tố ẩn như độ khó, cách trình bày nội dung, v.v. Thông qua việc phân rã, các yếu tố tiềm ẩn này được tạo ra, cho phép chúng ta thể hiện mối quan hệ phức tạp giữa sinh viên và môn học.

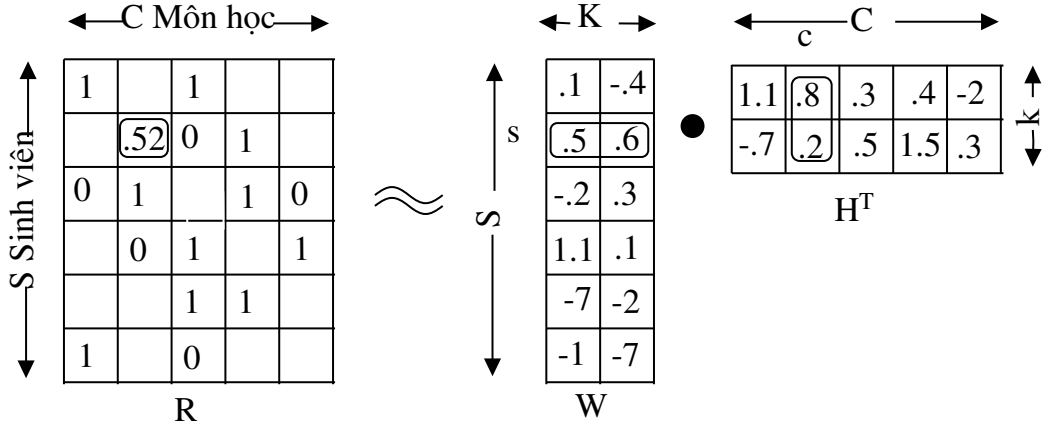
Quá trình phân rã ma trận thường được thực hiện thông qua các thuật toán tối ưu hóa. Mục tiêu của thuật toán là tìm ra các yếu tố tiềm ẩn mà khi nhân ma trận sinh viên-đặc điểm và ma trận môn học-đặc điểm lại, ta có thể tạo ra ma trận dự đoán gần với ma trận đánh giá ban đầu. Thuật toán sẽ cố gắng điều chỉnh các yếu tố tiềm ẩn này để sai số giữa dự đoán và thực tế là nhỏ nhất.

Phương pháp phân rã ma trận mang lại lợi ích trong việc giải quyết dữ liệu thưa và cung cấp cách tiếp cận linh hoạt để dự đoán kết quả học tập của sinh viên. Bằng cách xây dựng mối quan hệ giữa sinh viên và môn học thông qua các yếu tố tiềm ẩn, phương pháp này giúp hiểu sâu hơn về quá trình học tập và cải thiện độ chính xác của dự đoán trong môi trường giáo dục.

Chi tiết kỹ thuật phân rã ma trận trong dự đoán kết quả học tập của sinh viên được thực hiện như sau:

Kỹ thuật phân rã ma trận là việc chia một ma trận lớn G thành hai ma trận W và H , hai ma trận này có kích thước nhỏ hơn rất nhiều so với ma trận G , sao cho G có thể được

xây dựng lại từ hai ma trận nhỏ hơn này càng chính xác càng tốt [73] được minh họa trong Hình 3.3, nghĩa là $G \approx WH^T$.



Hình 3.3: Minh họa kỹ thuật phân rã ma trận

$W \in G^{|S| \times |K|}$ 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 $H \in G^{|C| \times |K|}$ 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 (3.15) như sau:

$$\hat{g}_{sc} = \sum_{k=1}^K w_{sk} h_{ck} = w_s \bullet h_c^T \quad (3.15)$$

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 theo điều kiện nào đó, chẳng hạn RMSE (Root Mean Squared Error) theo công thức (3.16) sau:

$$RMSE = \sqrt{\frac{1}{|D^{\text{test}}|} \sum_{s,c,g \in D^{\text{test}}} (g_{sc} - \hat{g}_{sc})^2} \quad (3.16)$$

Tối ưu hóa hàm mục tiêu theo phương pháp giảm dốc ngẫu nhiên (Stochastic Gradient Descent - SGD) [3] như công thức (3.17) sau:

$$o^{MF} = \sum_{(s,c,g) \in D^{\text{tan}}} \left(g_{sc} - \sum_{k=1}^K w_{sk} h_{ck} \right)^2 + \lambda (\|W\|_F^2 + \|H\|_F^2) \quad (3.17)$$

Với λ là hệ số chính tắc hóa ($0 \leq \lambda < 1$) và $\|\cdot\|_F^2$ là chuẩn Frobenius. Đại lượng $\lambda (\|W\|_F^2 + \|H\|_F^2)$ được 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 (3.18) và (3.19) sau:

$$w'_{sk} = w_{sk} + \beta (2e_{sc}h_{ck} - \lambda w_{sk}) \quad (3.18)$$

$$h'_{ck} = h_{ck} + \beta (2e_{sc}w_{sk} - \lambda h_{ck}) \quad (3.19)$$

Trong đó: $e_{sc} = g_{sc} - \hat{g}_{sc}$ 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 và biểu diễn như công thức (3.20) sau:

$$\hat{g}_{sc} = \sum_{k=1}^K w_{sk}h_{ck} \quad (3.20)$$

Thuật toán MF cho dự đoán kết quả học tập sinh viên

Chi tiết cho kỹ thuật phân rã ma trận trong bài toán dự đoán kết quả học tập của sinh viên được tóm tắt trong (thuật toán 3.1). Bản chất của thuật toán này dựa trên việc phân rã một ma trận lớn thành hai ma trận nhỏ hơn, một cho sinh viên và một cho các môn học, sao cho tích của chúng gần với ma trận ban đầu nhất.

1. Khởi tạo: Bắt đầu bằng việc khởi tạo hai ma trận đặc trưng W và H . W mô tả các đặc trưng ẩn của sinh viên và H mô tả đặc trưng ẩn của môn học. Cả hai ma trận này được khởi tạo với các giá trị ngẫu nhiên từ một phân phối chuẩn $N(0, 0.01)$.
2. Tính dự đoán và sai số: Với mỗi cặp sinh viên-môn học, thuật toán tính điểm dự đoán $\hat{g}_{sc}$ bằng cách nhân hàng tương ứng của W với cột tương ứng của H . Sai số giữa điểm thực tế và điểm dự đoán được xác định là $e_{sc} = g_{sc} - \hat{g}_{sc}$.
3. Cập nhật ma trận đặc trưng: Dựa trên sai số e_{sc} , các giá trị trong ma trận đặc trưng W và H được cập nhật. Để đảm bảo mô hình không overfitting và giảm sai số, thuật toán sử dụng hai tham số: hệ số học β để quy định tốc độ học và hệ số điều chuẩn λ để kiểm soát độ phức tạp của mô hình.

4. Lặp lại quá trình: Thuật toán tiếp tục cập nhật W và H cho đến khi một điều kiện dừng được thỏa mãn, chẳng hạn như sự giảm của sai số giữa hai lần lặp liên tiếp dưới mức ngưỡng $O_{Iter_{n-1}}^{MF} - O_{Iter_n}^{MF} < \varepsilon$, hoặc đạt đến một số lượng lặp tối đa.

Kết thúc quá trình này, thuật toán sẽ cung cấp hai ma trận đặc trưng W và H sao cho tích của chúng gần giống với ma trận điểm ban đầu, giúp dự đoán kết quả học tập của sinh viên trong các môn học mà họ chưa tham gia.

Giải thuật 3.1 Phân rã ma trận trong dự đoán kết quả học tập sinh viên

Data: D^{train} , K , beta, lamda, stopping condition

Result: W, H

```

1 Let  $s \in S$  be a student,  $c \in C$  be a course,  $g \in G$  be 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 while Stopping criterion is NOT met do
5   foreach  $(s, c, g_{sc})$  in  $D^{train}$  do
6      $\hat{g}_{sc} \leftarrow \sum_k^K (W[s][k] * H[c][k])$ 
7      $e_{sc} = g_{sc} - \hat{g}_{sc}$ 
8     for  $k \leftarrow 1$  to  $K$  do
9        $W[s][k] \leftarrow W[s][k] + \beta * (2e_{sc} * H[c][k] - \lambda * W[s][k])$ 
10       $H[c][k] \leftarrow H[c][k] + \beta * (2e_{sc} * W[s][k] - \lambda * H[c][k])$ 
11     end
12   end
13 end

```

Độ phức tạp thời gian của thuật toán 3.1 này phụ thuộc chủ yếu vào số lần lặp (iterations) và kích thước của ma trận đánh giá. Với số lượng sinh viên là s , số lượng môn học là c , số lượng lần lặp là i , và số lượng nhân tố tiềm ẩn là k (do k là hằng số rất nhỏ nên có thể bỏ qua). Vậy độ phức tạp thuật toán là $O(s \times c \times i)$

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

Đánh giá không đồng nhất (biased ratings) là một thách thức quan trọng trong các hệ thống gợi ý và trong việc xây dựng các mô hình dự đoán. Hiện tượng này có thể gây ảnh hưởng đáng kể đến chất lượng và hiệu suất của các mô hình. Dưới đây là một số cách để khắc phục vấn đề này:

1. Xử lý dữ liệu thiên vị: Trong quá trình tiền xử lý dữ liệu, bạn có thể áp dụng các kỹ

thuật để giảm thiểu tác động của thiên vị. Một cách phổ biến là chia tỷ lệ điểm đánh giá của mỗi người dùng hoặc đối tượng cho trung bình điểm đánh giá của họ. Điều này giúp đưa dữ liệu về một dạng tương đối không thiên vị.

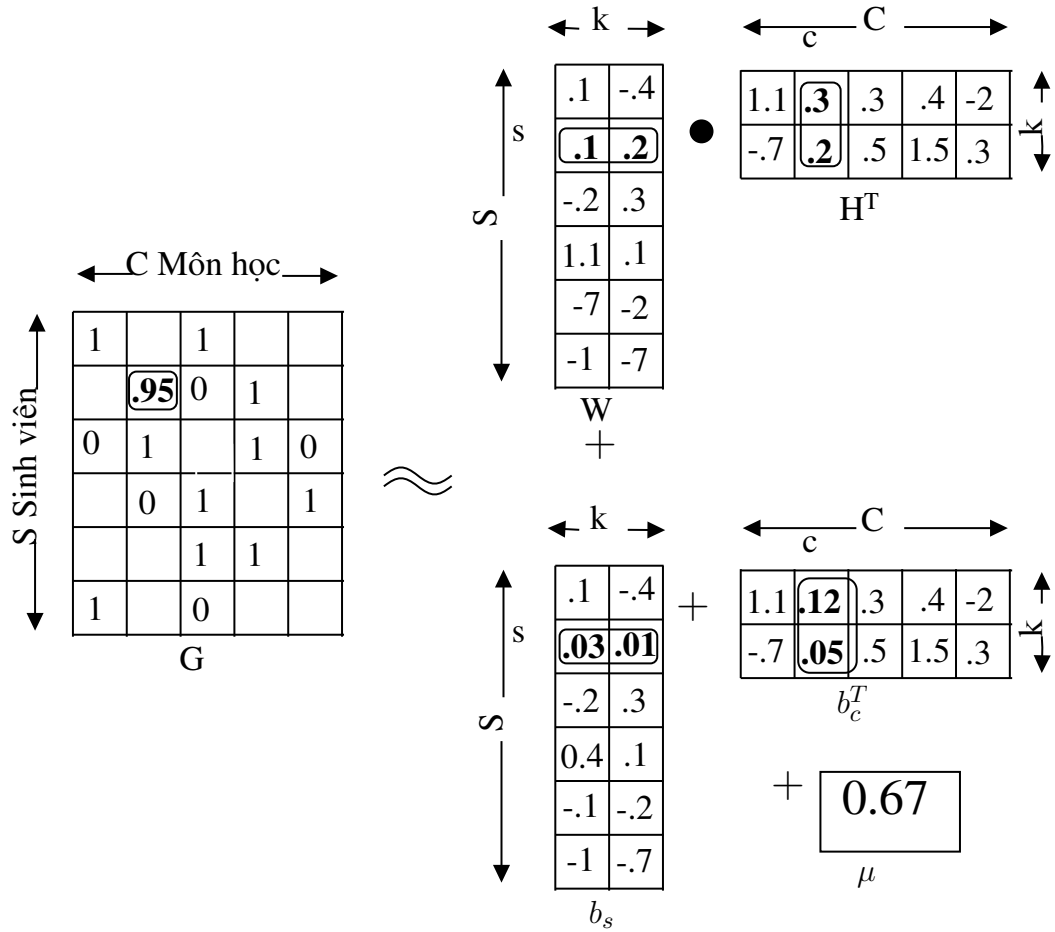
2. Mô hình hóa thiên vị: Bạn có thể xây dựng mô hình riêng để dự đoán và điều chỉnh thiên vị. Ví dụ, nếu một người dùng thường có xu hướng đánh giá cao, bạn có thể điều chỉnh các điểm đánh giá của họ xuống để tạo ra một dự đoán cân bằng hơn.
3. Kỹ thuật nhóm (grouping): Phân nhóm người dùng và đối tượng có thể giúp tạo ra các nhóm dữ liệu cân bằng hơn về mặt thiên vị. Điều này giúp các mô hình học từ các mẫu mang tính đại diện hơn về tất cả các khía cạnh của hệ thống.
4. Mô hình hóa tách biệt: Mô hình hóa riêng biệt cho các yếu tố khác nhau có thể giúp điều chỉnh thiên vị. Chẳng hạn, bạn có thể xây dựng các mô hình riêng cho người dùng và đối tượng, sau đó kết hợp chúng để tạo ra dự đoán cuối cùng.
5. Sử dụng đặc trưng thông tin bổ sung: Bổ sung thông tin như thông tin người dùng, thông tin đối tượng, hoặc các đặc trưng khác có thể giúp mô hình hiểu rõ hơn về ngữ cảnh và giảm thiểu tác động của thiên vị.

Trở lại bài toán dự đoán kết quả học tập sinh viên, Để giải quyết vấn đề dữ liệu khách quan nhằm đưa ra gợi ý chính xác nhất cho sinh viên. Luận án sử dụng kỹ thuật mô hình hóa thiên vị, nhằm 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. Sử dụng giải thuật Matrix Factorization kết hợp một lượng giá trị đo độ lệch Bias vào mô hình dự đoán (Xem 3.4), đề xuất được công bố trong công trình [CT1].

Dựa vào các khái niệm cơ sở của giải thuật Matrix Factorization (MF) và thêm giá trị lệch (bias) vào MF để được giải thuật Biased Matrix Factorization (Biased-MF) tránh tình trạng thiên vị trong đánh giá. Để dự đoán được năng lực của sinh viên s cho môn học c được biểu diễn với công thức (3.21) sau:

$$\hat{g}_{sc} = \mu + b_s + b_c + \sum_{k=1}^K \omega_{sk} h_{ck} \quad (3.21)$$

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 được tính theo công thức (3.22):



Hình 3.4: Minh họa kỹ thuật phân rã ma trận thưa

$$\mu = \frac{\sum (s, c, g) \in D^{\text{train}}}{|D^{\text{train}}|} \quad (3.22)$$

Giá trị b_s là độ lệch của sinh viên (giá trị lệch trung bình của năng lực một sinh viên so với giá trị trung bình toàn cục) được tính theo công thức (3.23):

$$b_s = \frac{\sum (s', c, g) \in D^{\text{train}} \mid s' = s(g - \mu)}{|\{(s', c, g) \in D^{\text{train}} \mid s' = s\}|} \quad (3.23)$$

Giá trị b_c là độ lệch của môn học (giá trị lệch trung bình của yêu cầu môn học so với giá trị trung bình toàn cục) được tính theo công thức (3.24) sau:

$$b_c = \frac{\sum (s, c', g) \in D^{\text{train}} \mid c' = c(g - \mu)}{|\{(s, c', g) \in D^{\text{train}} \mid c' = c\}|} \quad (3.24)$$

Do có thay đổi giá trị lệch của sinh viên và môn học, nên hàm mục tiêu cũng thay đổi

theo công thức (3.25) sau:

$$o^{BMF} = \sum_{(s,c,g) \in D^{rain}} \left(g_{sc} - \mu - b_s - b_c - \sum_{k=1}^K \omega_{sk} h_{ck} \right)^2 + \lambda (\|W\|_F^2 + \|H\|_F^2 + b_s^2 + b_c^2) \quad (3.25)$$

Cập nhật giá trị μ, b_c, b_s theo các công thức (3.26) (3.27) và (3.28) để cập nhật W và H sau tối ưu với hàm mục tiêu mới ta được:

$$\mu = \mu + \beta * e_{si} \quad (3.26)$$

$$b_s[s] \leftarrow b_s[s] + \beta * (e_{si} - \lambda * b_s[s]) \quad (3.27)$$

$$b_c[c] \leftarrow b_c[c] + \beta * (e_{sc} - \lambda * b_c[c]) \quad (3.28)$$

Cập nhật giá trị của w và h vẫn được tính toán theo hai công thức (3.18) và (3.19) của giải thuật MF cơ bản vì các giá trị tham số đã được cập nhật.

Sau quá trình huấn luyện ta được hai ma trận W và H đã tối ưu, thì quá trình dự đoán được thực hiện theo công thức (3.29) sau:

$$\hat{g}_{sc} = \mu + b_s + b_c + \sum_{k=1}^K \omega_{sk} h_{ck} \quad (3.29)$$

Thuật toán BiasedMF cho dự đoán kết quả học tập sinh viên

Mục tiêu của thuật toán BiasedMF là phân rã ma trận thành hai ma trận con (thể hiện đặc trưng của sinh viên và môn học) và thêm vào đó một số giá trị độ lệch (bias) để cải thiện độ chính xác của dự đoán.

1. Quá trình khởi tạo giá trị cho các đối tượng:

- Đầu vào là ma trận điểm D^{train} (sinh viên theo hàng, môn học theo cột).
- Khởi tạo hai ma trận đặc trưng W và H với các giá trị ngẫu nhiên tuân theo phân phối chuẩn $N(\mu, \sigma^2)$, với $\mu = 0$ và $\sigma^2 = 0.01$.
- Khởi tạo giá trị độ lệch cho từng sinh viên b_s và môn học b_c bằng 0.
- Tính μ là giá trị trung bình của tất cả các điểm trong D^{train} .

2. Quá trình cập nhật giá trị cho đến khi điều kiện dừng đạt được:

- Dự đoán điểm số $\hat{g}_{sc}$ của sinh viên s cho môn học c bằng tổng tích chập của

hàng sinh viên s trong W và hàng môn học c trong H .

- Tính sai số giữa điểm thực tế và dự đoán $e_{sc} = g_{sc} - \hat{g}_{sc}$.
- Cập nhật giá trị W và H dựa trên sai số e_{sc} , hệ số học β , và hệ số chính tắc λ để tránh quá khớp. Đồng thời, cập nhật giá trị độ lệch b_s và b_c dựa trên sai số e_{sc} .

Thuật toán này giúp tối ưu hóa việc dự đoán điểm số dựa trên đặc trưng của sinh viên và môn học, cũng như xem xét các yếu tố thiên vị trong việc đánh giá của từng sinh viên hoặc đặc điểm của từng môn học.

Giải thuật 3.2 Phân rã ma trận thiên vị trong dự đoán kết quả học tập sinh viên

Data: D^{train} , K , β , λ , stopping condition

Result: W, H

```

1 Let  $s \in S$  be a student,  $c \in C$  be a course,  $g \in G$  be 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 Initialize  $b_s$  and  $b_c$  to zero,  $\mu$  is the average of gradings
5 while Stopping criterion is NOT met do
6   foreach  $(s, c, g_{sc})$  in  $D^{train}$  do
7      $\hat{g}_{sc} \leftarrow \sum_k^K (W[s][k] * H[c][k])$ 
8      $e_{sc} = g_{sc} - \hat{g}_{sc}$ 
9     for  $k \leftarrow 1$  to  $K$  do
10       $W[s][k] \leftarrow W[s][k] + \beta * (2e_{sc} * H[c][k] - \lambda * W[s][k])$ 
11       $H[c][k] \leftarrow H[c][k] + \beta * (2e_{sc} * W[s][k] - \lambda * H[c][k])$ 
12       $b_s[s][k] \leftarrow b_s[s][k] + \beta * (2e_{sc} - \lambda * b_s[s][k])$ 
13       $b_c[c][k] \leftarrow b_c[c][k] + \beta * (2e_{sc} - \lambda * b_c[c][k])$ 
14    end
15  end
16 end

```

Độ phức tạp thời gian của thuật toán 3.2 này cũng phụ thuộc chủ yếu vào số lần lặp (iterations), kích thước của ma trận đánh giá và có bổ sung việc tính toán ma trận độ lệch. Với số lượng sinh viên là s , số lượng môn học là c , số lượng lần lặp là i , và số lượng nhân tố tiềm ẩn là k (do k là hằng số rất nhỏ nên có thể bỏ qua), và tương tự như vậy theo nguyên tắc lấy tối đa. Do đó độ phức tạp thuật toán vẫn là $O(s \times c \times i)$

3.6 Đánh giá kết quả

3.6.1 Tập dữ liệu

Tập dữ liệu điểm của khoa CNTT&TT, trường Đại học Cần Thơ (CTU) sau khi xử lý bao gồm 4017 sinh viên (4017 user) và 353 môn học (353 item) của 3 ngành học và gồm 279536 điểm chi tiết (279536 ratings).

Tập dữ liệu về điểm số của sinh viên từ trường Đại học Cần Thơ được coi là một nguồn thông tin đáng tin cậy cho việc xây dựng hệ thống gợi ý dự đoán kết quả học tập của sinh viên. Độ tin cậy của tập dữ liệu này thể hiện qua một số yếu tố chính sau:

Nguồn gốc dữ liệu: Dữ liệu được thu thập từ hệ thống quản lý điểm chính thức của trường. Hệ thống này thường có các biện pháp kiểm tra và xác thực để đảm bảo rằng dữ liệu nhập vào là chính xác. Điều này tăng cường độ tin cậy của tập dữ liệu.

Quy mô: Tập dữ liệu bao gồm 4017 sinh viên, 353 môn học và 279536 điểm chi tiết. Quy mô dữ liệu lớn này cho phép áp dụng các kỹ thuật phân tích dữ liệu tiên tiến, như lọc cộng tác, để tìm ra sự tương đồng giữa các sinh viên dựa trên lịch sử điểm số.

Chất lượng dữ liệu: Việc dữ liệu đã được xử lý để loại bỏ điểm nhiễu và điểm miễn môn học cho thấy nó đã trải qua quá trình tinh chỉnh để giảm thiểu sai sót và tăng cường độ chính xác.

Đa dạng: Tập dữ liệu bao gồm sinh viên từ 3 ngành học khác nhau, giúp đảm bảo độ đa dạng trong lịch sử học tập và cung cấp cơ hội phân tích sâu rộng về mô hình học tập giữa các ngành.

Ứng dụng cho hệ thống gợi ý: Dữ liệu thưa và dữ liệu lớn là một nguồn dữ liệu phù hợp để xây dựng hệ thống gợi ý. Với kỹ thuật lọc cộng tác, tập dữ liệu này có thể được sử dụng để tìm kiếm sự tương đồng giữa các sinh viên và dự đoán kết quả học tập của họ dựa trên lịch sử của sinh viên khác có kết quả học tập tương tự.

Tóm lại, tập dữ liệu điểm số của sinh viên từ trường Đại học Cần Thơ (CTU) có độ tin cậy cao và phù hợp để áp dụng trong việc xây dựng hệ thống gợi ý dự đoán kết quả học tập của sinh viên. Việc sử dụng dữ liệu này, kết hợp với các phương pháp phân tích dữ liệu tiên tiến, có tiềm năng cung cấp những gợi ý chính xác và hữu ích cho sinh viên.

3.6.2 Cài đặt thực nghiệm

Tìm kiếm siêu tham số

Tìm kiếm siêu tham số (hyper-parameter search) [74] để giải thuật đạt hiệu quả cao. Chúng tôi thực hiện việc tìm kiếm này qua hai giai đoạn: tìm thô (Coarse Search) và tìm mịn (Granularity Search) được thực hiện qua hai bước:

Bước 1: Tìm thô các giá trị tham số, ví dụ như k , β , λ . Do không gian tìm kiếm các tham số này khá lớn, nên việc tìm kiếm phải được chia ra nhiều phân đoạn, ta chỉ tìm trên các đầu mút phân đoạn.

Bước 2: Tìm mịn các giá trị tham số sau khi tìm được giá trị các tham số k , β , λ tốt nhất. Sau đó tìm lặp lại tìm lân cận của các giá trị tìm được, để tìm xem các giá trị lân cận của các tham số này có tốt hơn nữa hay không.

Ví dụ: sử dụng RMSE làm tiêu chí, kết quả tìm kiếm siêu tham số cho các mô hình trên tập dữ liệu CTU được trình bày trong Bảng 3.1.

Bảng 3.1: Các tham số thực nghiệm các giải thuật

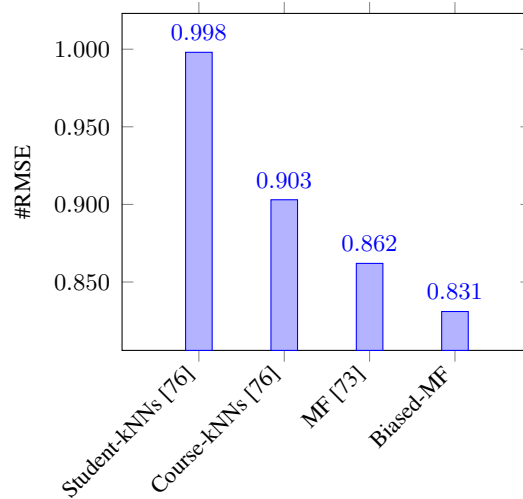
Phương pháp	Các siêu tham số	Thời gian	Độ phức tạp
Student-kNNs	$K = 5, \text{simMeasure} = \text{Cosine}$	7680s	$O(s^2 \times c)$
Course-kNNs	$K = 5, \text{simMeasure} = \text{Cosine}$	450s	$O(c^2 \times s)$
MF	$\beta = 0.01, \#iter = 30, K = 10, \lambda = 0.049$	10s	$O(s \times c \times i)$
Biased-MF	$\beta = 0.01, \#iter = 35, K = 10, \lambda = 0.42$	12s	$O(s \times c \times i)$

Cấu hình máy tính thực nghiệm

Để thực nghiệm các phương pháp, tác giả sử dụng máy tính có cấu hình: CPU Intel Core i7 và RAM 16GB. Các thuật toán cơ sở đã có được thực nghiệm dựa vào bộ thư viện CF4J [75], các thuật toán đề xuất do tác giả tự cài đặt và thực nghiệm theo cùng chiến lược thực nghiệm với phương pháp cơ sở.

3.6.3 Kết quả thử nghiệm

Trong bốn thí nghiệm được thực hiện, kết quả thử nghiệm được minh họa chi tiết tại Hình 3.5. Khi so sánh độ lỗi RMSE giữa các phương pháp, mô hình có sai số thấp nhất cho thấy nó là phương pháp hiệu quả nhất trong việc dự đoán kết quả học tập của sinh viên trong bộ dữ liệu được sử dụng.



Hình 3.5: Kết quả thực nghiệm trên tập dữ liệu CTU

Các phương pháp để so sánh dùng trong thực nghiệm

- **Student-kNNs:** là phương pháp lọc cộng tác theo bộ nhớ dựa vào sinh viên tương đồng, là biến thể của phương pháp User-based kNN, được thực nghiệm trên ngữ cảnh giáo dục. User-based kNN [76] là một phương pháp cơ sở trong hệ thống gợi ý, dựa trên sự tương đồng giữa người dùng. Bằng cách so sánh đánh giá của người dùng, nó dự đoán được giá trị đánh giá đối với các mục chưa được đánh giá. Mô hình này trong ngữ cảnh giáo dục giúp cá nhân hóa năng lực học tập của sinh viên và đề xuất môn học phù hợp. Tuy nhiên, phương pháp này phải đối mặt với thách thức từ dữ liệu thưa thớt và số lượng người dùng lớn.
- **Course-kNNs:** là phương pháp lọc cộng tác theo bộ nhớ dựa vào môn học tương đồng, là biến thể của phương pháp Item-based kNN được thực nghiệm trên ngữ cảnh giáo dục. Item-based kNN [76] là một phương pháp cơ sở trong hệ thống gợi ý. Phương pháp này so sánh các đánh giá của người dùng trên các đối tượng tương đồng với nhau, từ đó đưa ra dự đoán các giá trị đánh giá của những người dùng chưa đánh giá trên mục thông tin tương đồng. Mô hình này trong ngữ cảnh giáo dục cũng thể hiện dự đoán cá nhân hóa, và vẫn phải đối mặt với thách thức dữ liệu thưa và số lượng môn học lớn.
- **MF:** là phương pháp lọc cộng tác dựa vào mô hình phân rã ma trận, được thực nghiệm trên ngữ cảnh giáo dục. Mô hình phân rã ma trận [73] là một kỹ thuật học máy nhằm phân rã ma trận đánh giá thành hai ma trận con nhỏ. Mục tiêu chính của phương pháp này là tìm ra các ma trận con có kích thước nhỏ hơn sao cho khi nhân

chúng lại với nhau, ta thu được ma trận gần tiếp cận ma trận ban đầu. Thường được áp dụng trong hệ thống gợi ý, giúp khám phá những mối quan hệ ẩn giữa người dùng và mục tiêu, biểu diễn chúng dưới dạng các vector trong không gian ẩn.

- **Biased-MF**: Phương pháp lọc cộng tác dựa vào mô hình phân rã ma trận thiên vị trong ngữ cảnh giáo dục, nhằm giải quyết vấn đề đánh giá không đồng nhất, 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.

Đánh giá kết quả thực nghiệm

Phương pháp **Student-kNNs**, với RMSE đạt 0.998, cho thấy rằng việc dựa chỉ vào thông tin của người dùng trong môi trường giáo dục có thể không mang lại hiệu suất tốt nhất. Điều này nêu bật một hạn chế khi chỉ sử dụng thông tin cá nhân của sinh viên mà không xem xét các yếu tố khác.

Trái lại, **Course-kNNs** đã cải thiện đáng kể với RMSE là 0.903, nhấn mạnh rằng việc sử dụng thông tin về khóa học hoặc mục tiêu học tập có thể là một hướng đi tốt hơn khi tiếp cận thông tin người học. Tuy nhiên độ lỗi RMSE vẫn còn cao.

Khi tiếp tục nghiên cứu, phương pháp **MF** dựa trên mô hình phân rã ma trận đã giảm RMSE xuống 0.862, chứng minh rằng sử dụng lọc cộng tác dựa theo mô hình có thể cung cấp kết quả dự đoán chính xác hơn đối với tập dữ liệu lớn và thưa.

Tuy nhiên, điểm đáng chú ý nhất là **Biased-MF**, với RMSE chỉ **0.831**. Kết quả này chứng minh rằng việc tích hợp yếu tố thiên vị vào mô hình có thể giúp tăng cường độ chính xác của dự đoán, cung cấp một hướng đi mới cho việc cải thiện hệ thống dự đoán điểm số trong môi trường giáo dục.

Kết luận, việc áp dụng mô hình phân rã ma trận, đặc biệt khi tích hợp yếu tố thiên vị, đã chứng tỏ sự ưu việt trong việc dự đoán kết quả học tập so với những phương pháp khác.

3.7 Tổng kết chương

Chương này đã trình bày và đánh giá các phương pháp khác nhau trong việc dự đoán kết quả học tập của sinh viên. Qua quá trình thử nghiệm và so sánh, rõ ràng là việc tích

hợp yếu tố thiên vị vào mô hình phân rã ma trận đã đem lại hiệu quả đáng chú ý, nổi bật trước các phương pháp khác. Những nhận định từ kết quả này không chỉ làm rõ ưu và khuyết điểm của mỗi phương pháp, mà còn đề xuất hướng đi mới trong việc nâng cao chất lượng hệ thống gợi ý trong lĩnh vực giáo dục.

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 mô hình 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

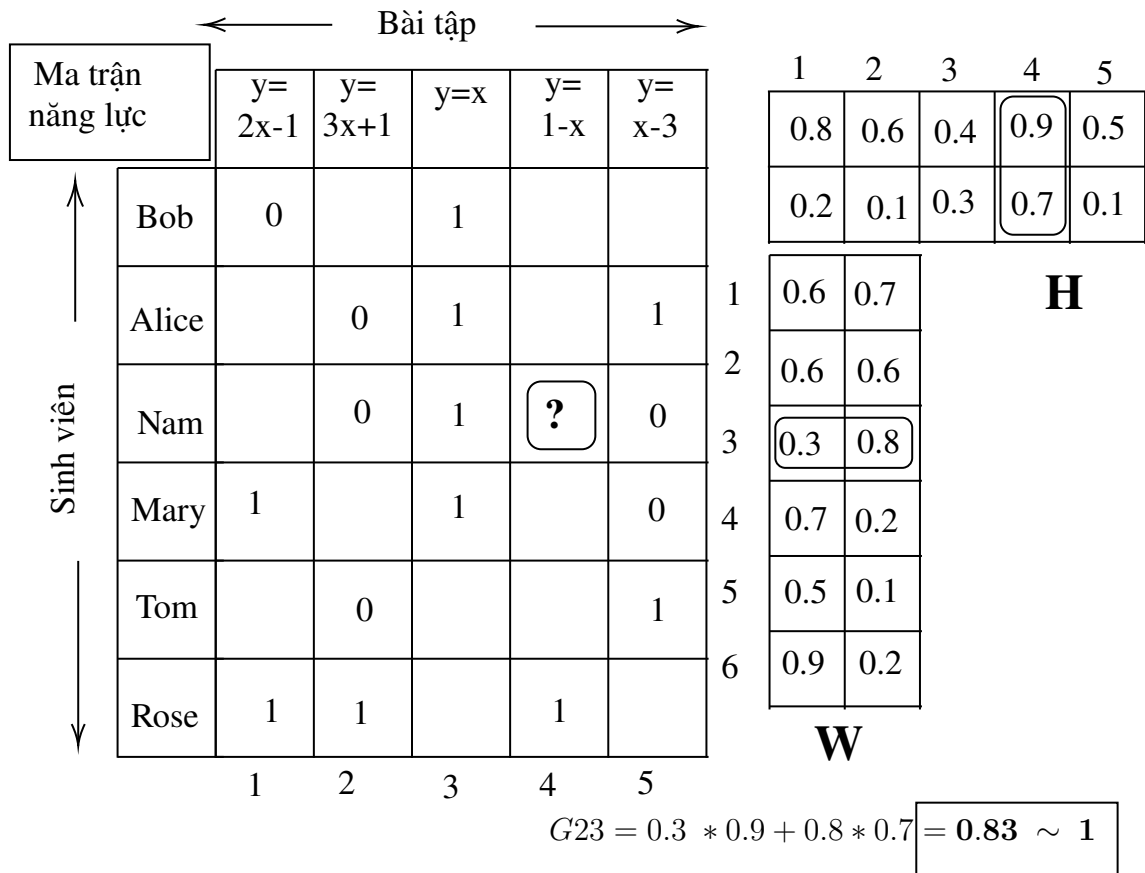
Nội dung chương này là trình bày kết quả nghiên cứu của luận án về việc đề xuất tích hợp kiến trúc học sâu vào kỹ thuật phân rã ma trận để dự đoán kết quả học tập của sinh viên. Kết quả thực nghiệm của đề xuất này đã nâng cao được độ chính xác cho dự đoán và hoàn toàn thích hợp để xây ứng dụng vào hệ thống gợi ý môn học trong việc cố vấn học tập. Bên cạnh đó, ở chương này cũng trình bày thêm phương pháp mở rộng là phân rã sâu ma trận có kết hợp giá trị độ lệch làm tham số đầu vào cho mô hình dự đoán kết quả học tập của sinh viên nhằm cải thiện thêm độ chính xác cho dự đoán. Nội dung này được tổng hợp từ công trình nghiên cứu đã công bố [CT4]

4.1 Bài toán nghiên cứu

Như được giới thiệu bài toán nghiên cứu đã trình bày ở phần 1.2 của luận án. Ở phần này, luận án xin được trình bày nhanh về bài toán dự đoán kết quả học tập của sinh viên theo hướng tiếp cận kỹ thuật phân rã ma trận (Matrix factorization - MF), từ đó triển khai các mô hình cải tiến như: Phân rã ma trận thiên vị (Biased Matrix Factorization - Biased-MF) để giải quyết vấn đề đánh giá không đồng nhất, cũng như mô hình tích hợp kiến trúc học sâu vào kỹ thuật phân rã ma trận (Deep Learning Matrix Factorization - DMF) nhằm nâng cao độ chính xác cho dự đoán kết quả học tập của sinh viên, để hỗ trợ gợi ý lựa chọn môn học trong việc cố vấn học tập.

Bài toán dự đoán kết quả học tập của sinh viên được chuyển sang mô hình phân rã ma trận của hệ thống gợi ý được trình bày như (Hình 4.1). Giả sử có 6 sinh viên và 5 bài tập được trình bày trên một ma trận X . Ví dụ, bài tập của sinh viên là tính giá trị y từ giá trị x theo hàm $y = 1 - x$. Năng lực của sinh viên được đánh giá với thang điểm là 1 (đúng) hoặc 0 (sai). Vấn đề của chúng ta là dự đoán những bài tập mà sinh viên chưa làm (những ô trống trong ma trận X). Sau khi có kết quả dự đoán của tất cả ô trống, ta có thể đưa ra gợi ý cụ thể trên những ràng buộc tình huống, chẳng hạn như ràng buộc giới

hạn số lượng các môn học, bài tập cần cải thiện và luyện tập.



Hình 4.1: Ví dụ về mô hình dự đoán năng lực của sinh viên khi làm bài tập

Để triển khai mô hình dự đoán, hệ thống gợi ý để dự đoán kết quả học tập của sinh viên sử dụng kỹ thuật phân rã ma trận. Ở đó, mỗi sinh viên được xem như là người dùng (user), mỗi môn học được xem như là mục thông tin (item), và kết quả học tập được xem như là đánh giá (rating) trong hệ thống gợi ý. Tuy vậy, khác với những nghiên cứu trước là dự đoán kết quả cho cả học kỳ (điểm GPA), trong khuôn khổ nghiên cứu này, luận án thực hiện việc dự đoán kết quả cho từng môn học và chỉ dựa trên thông tin tương tác (collaboration) giữa sinh viên và môn học mà không dùng đến các thông tin nhân khẩu học (mặc dù vậy, kỹ thuật này hoàn toàn có thể dùng để dự đoán kết quả cho từng học kỳ như những nghiên cứu trước đây).

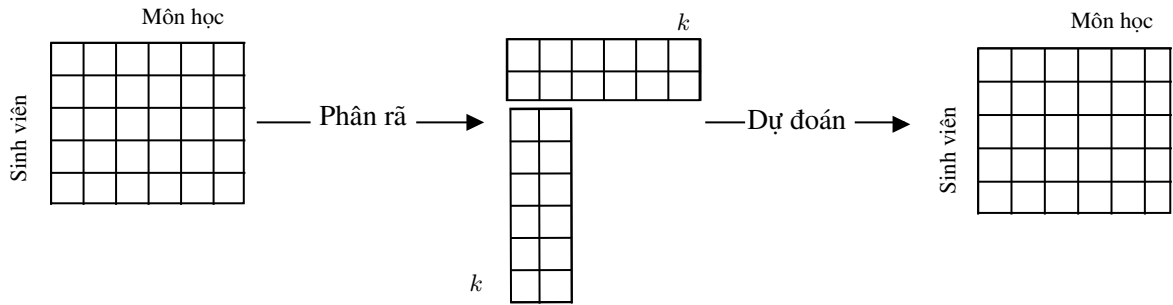
Trong chương này, luận án đề xuất việc áp dụng kiến trúc học sâu vào kỹ thuật phân rã ma trận nhằm nâng cao kết quả dự đoán. Nội dung chi tiết được trình bày trong phần 3.3, và kết quả của nghiên cứu đã được công bố trong [CT4].

4.2 Kỹ thuật phân rã sâu ma trận trong bài toán dự đoán kết quả học tập của sinh viên (Deep Matrix Factorization for Student Performance Prediction)

4.2.1 Phương pháp phân rã ma trận

Bài toán kỹ thuật phân rã ma trận trong dự đoán kết quả học tập của sinh viên đã được trình bày chi tiết ở phần 3.4. Tuy nhiên, trong phần này chúng tôi xin trình bày phương pháp dưới dạng khác, theo định dạng để có thể cải tiến mô hình phân rã ma trận thành phân rã sâu, phân rã qua nhiều tầng của ma trận để đạt được độ chính xác cao hơn.

Kỹ thuật phân rã ma trận gồm hai giai đoạn: Đầu tiên là phân rã (Factorization) một ma trận lớn G thành hai ma trận W và H , hai ma trận này có kích thước nhỏ hơn rất nhiều so với ma trận G ; Sau đó, G có thể được xây dựng lại hay còn gọi là dự đoán (Prediction) từ hai ma trận nhỏ hơn này càng chính xác càng tốt, nghĩa là $G \approx WH^T$ [73]. Hình 4.2 mô tả chi tiết hai giai đoạn của kỹ thuật phân rã ma trận trong bài toán dự đoán kết quả học tập của sinh viên. Đây cũng là cơ sở cho việc cải tiến mô hình phân rã sâu ma trận ở phần tiếp theo.



Hình 4.2: Minh họa kỹ thuật phân rã ma trận

Quá trình phân rã (Factorization): Khởi tạo 2 ma trận con ngẫu nhiên, sau đó cập nhật các giá trị của hai ma trận này cho đến khi điều kiện dừng thỏa mãn.

Ma trận $W \in G^{|S| \times |K|}$ là một ma trận đại diện cho sinh viên 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 Ma trận $H \in G^{|C| \times |K|}$ là một ma trận đại diện cho môn học mà ở đó mỗi dòng c là một véc-tơ bao gồm k nhân tố tiềm ẩn mô tả cho môn học c . Khi đó điểm số g của sinh viên s trên môn học c được dự đoán bởi công thức (4.1) như sau:

$$\hat{g}_{sc} = \sum_{k=1}^K w_{sk} h_{ck} = w_s \bullet h_c^T \quad (4.1)$$

Hàm mục tiêu RMSE (Root Mean Squared Error) được tối ưu hóa theo phương pháp giảm dốc ngẫu nhiên (Stochastic Gradient Descent - SGD) [3] như công thức (4.2) sau:

$$o^{MF} = \sum_{(s,c,g) \in D^{\text{tan}}} \left(g_{sc} - \sum_{k=1}^K w_{sk} h_{ck} \right)^2 + \lambda (\|W\|_F^2 + \|H\|_F^2) \quad (4.2)$$

Với $(0 \leq \lambda < 1)$ là hệ số chính tắc hóa; $\|\cdot\|_F^2$ là chuẩn Frobenius. Đại lượng $\lambda (\|W\|_F^2 + \|H\|_F^2)$ dùng để ngăn ngừa sự quá khớp (over-fitting). Công thức (4.3)(4.4) cập nhật lại các giá trị w và h (Với $e_{sc} = g_{sc} - \hat{g}_{sc}$ là độ lỗi của dự đoán, β là tốc độ học):

$$w'_{sk} = w_{sk} + \beta (2e_{sc} h_{ck} - \lambda w_{sk}) \quad (4.3)$$

$$h'_{ck} = h_{ck} + \beta (2e_{sc} w_{sk} - \lambda h_{ck}) \quad (4.4)$$

Quá trình dự đoán (Prediction): Sau khi quá trình huấn luyện ta được hai ma trận W và H đã tối ưu, thì quá trình dự đoán được thực hiện và tính như công thức (4.5) sau:

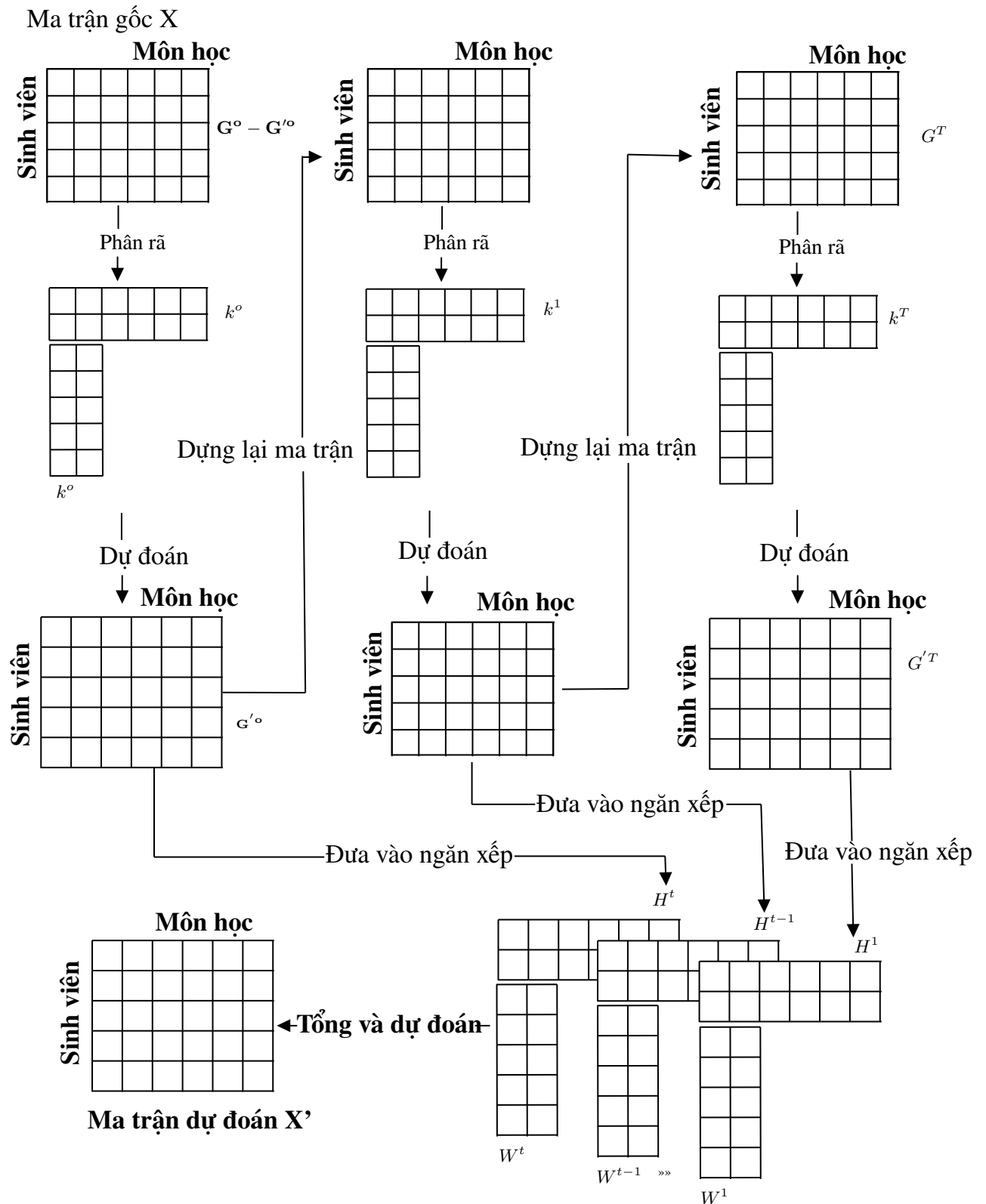
$$\hat{g}_{sc} = \sum_{k=1}^K w_{sk} h_{ck} \quad (4.5)$$

4.2.2 Phương pháp phân rã sâu 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 đáp ứng kết quả tốt nhất. Mỗi bước trong quy trình là một phương pháp cơ sở (MF); mô hình hoàn chỉnh là tổng các ma trận được lưu trên ngăn xếp của các bước thực hiện phân rã.

Phương pháp tích hợp thực hiện việc phân rã ma trận qua nhiều tầng để đạt được kết quả tối ưu nhất. Mô hình tích hợp này sử dụng các nguyên tắc của học sâu để tinh chỉnh ma trận đầu ra của mô hình thông qua huấn luyện mô hình liên tiếp. Phương pháp này

được gọi là phân rã sâu ma trận (DMF). Hình 4.3 minh họa hoạt động của DMF. Ta có thể quan sát mô hình được khởi tạo với đầu vào là các tham số cơ bản của RS dựa trên lọc cộng tác: một ma trận điểm số X của sinh viên đối với môn học hoặc bài tập nào đó.



Hình 4.3: Mô hình tích hợp kiến trúc học sâu vào phân rã ma trận

Như trong phương pháp MF, ma trận X này cũng sẽ được gọi là $X = G^0$, 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 chúng 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 tích thừa số 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 tích nhân tử các ma trận lỗi liên tiếp $G^1, \dots, G^T$. Có lẽ, chuỗi ma trận lỗi này hội tụ về 0, vì vậy chúng tôi nhận được các dự đoán trước khi chúng tôi thêm các lớp mới vào mô hình.

Tương tự như phương pháp MF chuẩn đã được trình bày ở trên. Phương pháp này có thể được chính thức hóa như sau. Dự đoán kết quả học tập của học sinh là xử lý n học sinh và m khóa học, có điểm được thu thập trong một ma trận dự phòng $R = (R_{(s,c)}) \in R_{(n \times m)}$ trong đó $R_{(s,c)}$ là điểm mà sinh viên đã học khóa học c (thông thường, các giá trị float từ 0 đến 1), hoặc $R_{(s,c)} = \emptyset$ nếu sinh viên chưa học khóa c . Chúng tôi tìm kiếm thừa số hóa $G^0 = R$ có dạng $G^0 \approx W^0 \cdot H^0$, trong đó W^0 là ma trận $n \times k^0$ và H^0 là ma trận $k^0 \times m$. Ở đây, k^0 được hiểu là số lượng các yếu tố ẩn mà chúng tôi cố gắng phát hiện trong bước đầu tiên (thường k^0 là khoảng từ 5 đến 9) và W^0, H^0 được xem là phép chiếu / đồng chiếu của n sinh viên và khóa học m vào không gian tiềm ẩn k^0 chiều. Các ma trận xếp hạng nhỏ này được học sao cho tích $W^0 \cdot H^0$ là một xấp xỉ tốt của ma trận xếp hạng $G^0 = R$, nghĩa là theo khoảng cách Euclide thông thường theo công thức (4.6) sau:

$$G^0 \approx \hat{G}^0 = W^0 \cdot H^0 \quad (4.6)$$

Để triển khai mô hình học sâu, chúng tôi trừ xấp xỉ được thực hiện bởi $\hat{G}^0$ cho ma trận ban đầu X , để thu được ma trận thừa thớt mới G^1 có chứa lỗi dự đoán ở lần lặp đầu tiên như công thức (4.7):

$$G^1 = R - \hat{G}^0 = G^0 - W^0 \cdot H^0 \quad (4.7)$$

Lưu ý 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. Thật vậy, sự điều chỉnh này là ý tưởng chính của việc áp dụng kiến trúc học sâu. Để làm điều này, ta cần thực hiện một phép phân tích nhân tử mới cho ma trận lỗi G^1 theo công thức (4.8) sau:

$$G^1 \approx \widehat{G}^1 = W^1 \cdot H^1 \quad (4.8)$$

Quá trình xấp xỉ trong mỗi bước được thực hiện tương tự. Tuy nhiên, hai ma trận W^1, H^1 có thứ tự $s \times k^1$ và $k^1 \times c$ cho một số xác định của hệ số tiềm ẩn k^1 . Lưu ý rằng 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 bởi công thức (4.9):

$$G^t = G^{t-1} - \widehat{G}^{t-1} = G^{t-1} - W^{t-1} \cdot H^{t-1} \quad (4.9)$$

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 theo công thức (4.10) sau:

$$G^t \approx \widehat{G}^t = W^t \cdot H^t \quad (4.10)$$

Khi quá trình phân tích nhân tử 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 như công thức tổng quát (4.11) sau:

$$\begin{aligned} R \approx \widehat{R} &= \widehat{G}^0 + G^1 = \widehat{G}^0 + \widehat{G}^1 + G^2 = \dots = \widehat{G}^0 + \widehat{G}^1 + \widehat{G}^2 + \dots + \widehat{G}^T \\ &= W^0 \cdot H^0 + W^1 \cdot H^1 + \dots + W^T \cdot H^T = \sum_{t=0}^T W^t \cdot H^t \end{aligned} \quad (4.11)$$

Đối với bất kỳ bước $t = 0, \dots, T$ nào, sự phân tích nhân tử $G^t \approx W^t \cdot H^t$ được tìm

theo phương pháp tiêu chuẩn để giảm thiểu khoảng cách euclide giữa G^t và $W^t \cdot H^t$ theo đường giảm dần với chính quy. Hàm lỗi cho DMF bây giờ được tính như công thức (4.12) như sau:

$$O^{DLMF} = \sum_{t=0}^T \left(\sum_{(s,c,g) \in D^t \text{ train}} \left(g_{sc}^t - \sum_{k=1}^K w_{sk}^t h_{kc}^t \right)^2 + \lambda^t \left(\|W^t\|_F^2 + \|H^t\|_F^2 \right) \right) \quad (4.12)$$

Trong đó thuật ngữ λ^t là siêu tham số chính tắc của bước t để tránh việc học vẹt. Hàm lỗi O^{DMF} có thể được dẫn xuất thành w_s^t và h_c^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 hai công thức (4.13) và (4.14) như sau:

$$w_{sk}^{t'} = w_{sk}^t + \beta (2e_{sc} h_{ck}^t - \lambda w_{sk}^t) \quad (4.13)$$

$$h_{ck}^{t'} = h_{ck}^t + \beta (2e_{sc} w_{sk}^t - \lambda h_{ck}^t) \quad (4.14)$$

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

Theo cách này, sau khi kết thúc quá trình phân rã 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 (4.15):

$$\hat{g}_{sc} = \sum_{t=0}^T \sum_{k=1}^K w_{sk}^t h_{kc}^t = \sum_t w_s^t * (h_c^t)^T \quad (4.15)$$

Lưu ý rằng phương pháp này bao gồm các lần lặp lại liên tiếp của một quá trình MF sử dụng kết quả của MF trước đó làm đầu vào. Tất cả các siêu tham số được lưu trữ trong ngăn xếp, vì vậy chúng tôi có thể dễ dàng sử dụng cách tiếp cận đệ quy và thuật toán triển khai đệ quy.

Thuật toán Predicting Student Performance-DMF

Thuật toán nhận đầu vào là ma trận gốc X và các siêu tham số của mô hình (một ngăn xếp chứa tập các tham số của quá trình phân rã ma trận). Tương tự, đầu ra của thuật toán sẽ là một ngăn xếp chứa các cặp (W, H) thể hiện đầy đủ khả năng dự đoán của quá trình học sâu.

Lưu ý rằng các siêu tham số này được xếp chồng lên nhau để mỗi quá trình phân rã ma trận được thực hiện sử dụng các tham số khác nhau. Các siêu tham số của quá trình phân rã ma trận đầu tiên sẽ được đẩy lên trên cùng của ngăn xếp và các tham số của quá trình phân rã ma trận thứ hai ở vị trí tiếp theo. Và cứ tiếp tục như vậy cho đến khi các tham số của lần phân rã số cuối cùng sẽ được đẩy xuống dưới cùng của ngăn xếp. Điều này cho phép chúng tôi xác định tiêu chí dừng của thuật toán là độ sâu của ngăn xếp, thường là khoảng bốn tầng hay khi các độ lỗi được tối ưu ($O_{Iter_{t-1}}^{DMF} - O_{Iter_t}^{DMF} < \varepsilon$).

Chi tiết về phương pháp đề xuất được trình bày trong thuật toán 4.1 (Predicting Student Performance-DMF). Thuật toán này thực hiện tinh chỉnh phân rã ma trận qua các lần gọi đệ quy bằng cách sử dụng các ngăn xếp tham số như: k hệ số tiềm ẩn, tốc độ học β , trọng lượng chính quy λ , điều kiện dừng và độ sâu.

Ví dụ, chúng tôi sử dụng các siêu tham số để thực hiện khối các câu lệnh MF ở mỗi độ sâu. Trong mỗi khối câu lệnh MF ở dòng 1-15 được thực hiện như phương pháp MF tiêu chuẩn (được trình bày ở phần 3.4): đầu tiên là khởi tạo ngẫu nhiên ở dòng 1-4, kế tiếp quá trình huấn luyện hai ma trận w và h ở dòng lệnh 10-11. Sau đó, tinh chỉnh ma trận dự đoán so với ma trận đầu vào ở bước lặp này (dòng 19) để đẩy các mô hình đào tạo hoàn chỉnh vào ngăn xếp ở (dòng 20-21). Thuật toán nhận đầu vào là ma trận R , chứa các điểm của sinh viên đối với các môn học ($R = G^0$) cho lần gọi đầu tiên hoặc các lỗi của lần phân rã ma trận trước đó G^t cho các lần gọi kế tiếp.

Giải thuật 4.1 PredictingStudentPerformance-DMF

Data: D^{train} , K , β s, λ s, stopping condition, depth

Result: W, H

```

1 Let  $s \in S$  be a student,  $c \in C$  be a course,  $g \in G$  be 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   foreach  $(s, c, g_{sc})$  in  $D^{train}$  do
7      $\hat{g}_{sc} \leftarrow \sum_k^K (W[s][k] * H[c][k])$ 
8      $e_{sc} = g_{sc} - \hat{g}_{sc}$ 
9     for  $k \leftarrow 1$  to  $K$  do
10       $W[s][k] \leftarrow W[s][k] + \beta * (2e_{sc} * H[c][k] - \lambda * W[s][k])$ 
11       $H[c][k] \leftarrow H[c][k] + \beta * (2e_{sc} * W[s][k] - \lambda * H[c][k])$ 
12    end
13  end
14 end
15 if (is-empty( $K$ )) then
16   return new stack ( $\langle W, H \rangle$ )
17 else
18    $\hat{G} = G - W \cdot H$ 
19   Params-Stack = Deep-Learning-Matrix-Factorization( $\hat{G}, K, T, \text{Betas}, \text{Lamdas}$ )
20   Return push( $\langle W, H \rangle$ , Params-Stack)
21 end

```

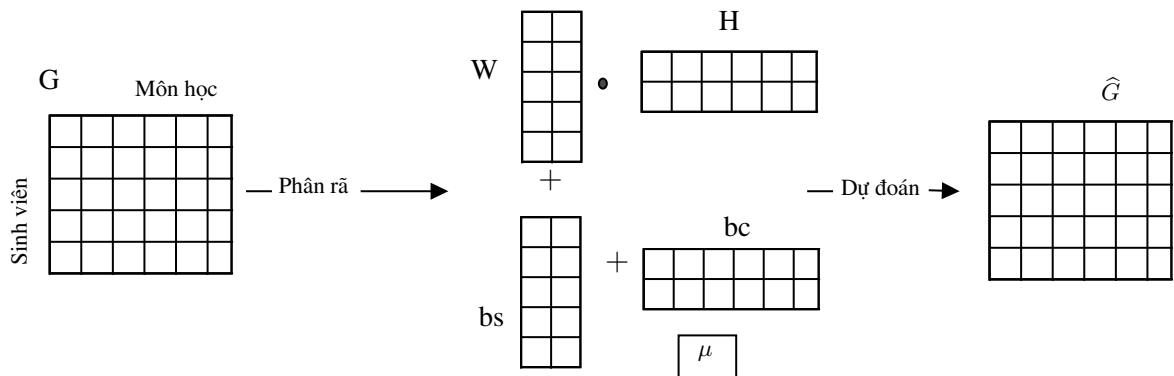
Độ phức tạp thời gian của thuật toán 4.1 này cũng phụ thuộc chủ yếu vào số lần lặp (iterations), kích thước của ma trận đánh giá và có bổ sung số lần phân rã sâu. Với số lượng sinh viên là s , số lượng môn học là c , số lượng lần lặp là i , số tầng phân rã sâu là d và số lượng nhân tố tiềm ẩn là k (do k là hằng số rất nhỏ nên có thể bỏ qua), và tương tự như vậy theo nguyên tắc lấy tối đa. Do đó độ phức tạp thuật toán là $O(s \times c \times i \times d)$

4.3 Kỹ thuật phân rã sâu ma trận thiên vị trong dự đoán kết quả học tập sinh viên (Deep Biased Matrix Factorization for Student Performance Prediction)

4.3.1 Phương pháp phân rã ma trận thiên vị

Bài toán kỹ thuật phân rã ma trận thiên vị trong dự đoán kết quả học tập của sinh viên đã được trình bày chi tiết ở phần 3.5. Tuy nhiên, trong phần này chúng tôi xin trình bày phương pháp theo định dạng có thể cải tiến mô hình phân rã ma trận thành phân rã sâu, phân rã qua nhiều tầng của ma trận để đạt được độ chính xác cao hơn.

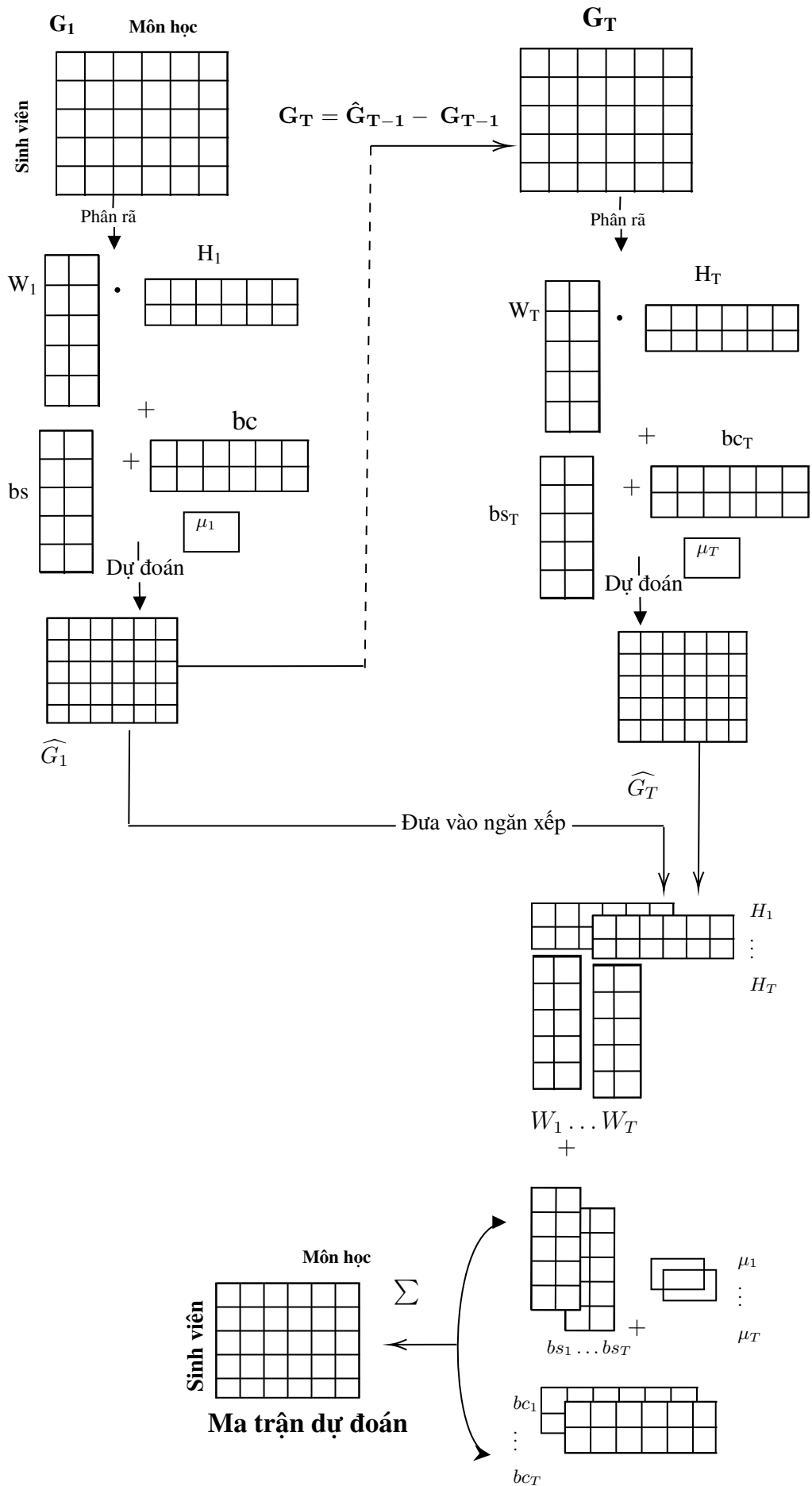
Để giải quyết vấn đề dữ liệu khách quan nhằm đưa ra gợi ý chính xác nhất cho sinh viên. Giảm bớt sự chênh lệch giữa những yêu cầu cao thấp khác nhau của các môn học. Cũng như giảm thiểu sự gợi ý sai lệch do nhìn nhận từ những sinh viên có sở trường hay sở đoản đối với môn học nào đó. Giải thuật Matrix Factorization được sử dụng để kết hợp một lượng giá trị đo độ lệch Bias vào mô hình dự đoán, phương pháp được công bố trong công trình [CT1] (Xem Hình 4.4).



Hình 4.4: Minh họa kỹ thuật phân rã ma trận thiên vị

4.3.2 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 mới 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). Hình 4.5 minh họa hoạt động của DBMF với khởi đầu là một ma trận điểm G của sinh viên với môn học.



Hình 4.5: Mô hình tích hợp kiến trúc học sâu vào phân rã ma trận thiên vị

Thuật toán Predicting Student Performance-DBMF

Thuật toán phân rã ma trận thiên vị cho bài toán dự đoán kết quả học tập của sinh viên nhận đầu vào là ma trận gốc X và các siêu tham số của mô hình (một ngăn xếp chứa tập các tham số của quá trình phân rã ma trận). Tương tự, đầu ra của thuật toán sẽ là một ngăn xếp chứa các cặp (W, H) thể hiện đầy đủ khả năng dự đoán của quá trình học sâu. Thuật toán này tương tự giải thuật DMF của phần 4.2, tuy nhiên chúng được mở rộng với các giá trị độ lệch Bias của sinh viên và của môn học.

Lưu ý rằng các siêu tham số này được xếp chồng lên nhau để mỗi quá trình phân rã ma trận được thực hiện sử dụng các tham số khác nhau. Các siêu tham số của quá trình phân rã ma trận đầu tiên sẽ được đẩy lên trên cùng của ngăn xếp và các tham số của quá trình phân rã ma trận thứ hai ở vị trí tiếp theo. Và cứ tiếp tục như vậy cho đến khi các tham số của lần phân rã số cuối cùng sẽ được đẩy xuống dưới cùng của ngăn xếp. Điều này cho phép chúng tôi xác định tiêu chí dừng của thuật toán là độ sâu của ngăn xếp, thường là khoảng bốn tầng hay khi các độ lỗi được tối ưu ($O_{Iter_{t-1}}^{DBMF} - O_{Iter_t}^{DBMF} < \varepsilon$).

Chi tiết về phương pháp đề xuất được trình bày trong thuật toán 4.2 (Predicting Student Performance-DMF). Thuật toán này thực hiện tinh chỉnh phân rã ma trận qua các lần gọi đệ quy bằng cách sử dụng các ngăn xếp tham số như: k hệ số tiềm ẩn, tốc độ học β , trọng lượng chính quy λ , điều kiện dừng và độ sâu.

Ví dụ, chúng tôi sử dụng các siêu tham số khác nhau để thực hiện khối các câu lệnh Biased-MF ở mỗi độ sâu khác nhau. Trong mỗi khối câu lệnh Biased-MF ở dòng (1-17) được thực hiện như phương pháp Biased-MF tiêu chuẩn (được trình bày ở phần 3.5) bao gồm: đầu tiên là khởi tập các giá trị ngẫu nhiên (dòng 1-5), tiếp theo là quá trình huấn luyện (cập nhật các giá trị w h ở (dòng 11-12) cập nhật các giá trị độ lệch Bias (dòng 13-14)). Sau đó, tinh chỉnh ma trận (dòng 21) để đẩy các mô hình đào tạo hoàn chỉnh vào ngăn xếp ở (dòng 22-23). Thuật toán nhận đầu vào là ma trận R , chứa các điểm của sinh viên đối với các môn học ($R = G^0$) cho lần gọi đầu tiên hoặc các lỗi của lần phân rã ma trận trước đó G^t cho các lần gọi kế tiếp.

Giải thuật 4.2 PredictingStudentPerformance-DBMF

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

Result: W, H

```

1 Let  $s \in S$  be a student,  $c \in C$  be a course,  $g \in G$  be 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 Initilize  $b_s$  and  $b_c$  to zero,  $\mu$  is the average of gradings
6 while Stopping criterion is NOT met do
7   foreach  $(s, c, g_{sc})$  in  $D^{train}$  do
8      $\hat{g}_{sc} \leftarrow \sum_k^K (W[s][k] * H[c][k])$ 
9      $e_{sc} = g_{sc} - \hat{g}_{sc}$ 
10    for  $k \leftarrow 1$  to  $K$  do
11       $W[s][k] \leftarrow W[s][k] + \beta * (2e_{sc} * H[c][k] - \lambda * W[s][k])$ 
12       $H[c][k] \leftarrow H[c][k] + \beta * (2e_{sc} * W[s][k] - \lambda * H[c][k])$ 
13       $b_s[s][k] \leftarrow b_s[s][k] + \beta * (2e_{sc} - \lambda * b_s[s][k])$ 
14       $b_c[c][k] \leftarrow b_c[c][k] + \beta * (2e_{sc} - \lambda * b_c[c][k])$ 
15    end
16  end
17 end
18 if  $(\text{is-empty}(K))$  then
19   return new stack ( $< W, H, b_s, b_c, \mu >$ )
20 else
21    $\hat{G} = G - (W \cdot H + b_s + b_c + \mu)$ 
22   Params-Stack = Deep-Biased-Matrix-Factoriztion( $\hat{G}, K, T, \text{Betas}, \text{Lamdas}$ )
23   Return push( $< W, H, b_s, b_c, \mu >$ , Params-Stack)
24 end

```

Độ phức tạp thời gian của thuật toán 4.2 này cũng phụ thuộc chủ yếu vào số lần lặp (iterations), kích thước của ma trận đánh giá có bổ sung thêm độ lệch nhưng là hằng số không đáng kể, và số lần phân rã sâu. Với số lượng sinh viên là s , số lượng môn học là

c , số lượng lần lặp là i , số tầng phân rã sâu là d và số lượng nhân tố tiềm ẩn là k (do k là hằng số rất nhỏ nên có thể bỏ qua), và tương tự như vậy theo nguyên tắc lấy tối đa. Do đó độ phức tạp thuật toán vẫn là $O(s \times c \times i \times d)$

4.4 Đánh giá kết quả

4.4.1 Tập dữ liệu

Tập dữ liệu ASSISTments được công bố dành cho cộng đồng nghiên cứu về khai thác dữ liệu giáo dục [70]. Tập dữ liệu được phát hành bởi ASSISTments Platform, là hệ thống trợ giảng thông minh trên nền web, giúp giáo viên đánh giá quá trình học học môn toán của sinh viên. Hệ thống cho phép giáo viên viết những bài tập riêng lẻ, mỗi bài tập bao gồm câu hỏi, gợi ý liên quan, video hướng dẫn, đáp án v.v. . . Cấu trúc dạng phân cấp được biểu diễn như sau: $Assignment \supseteq Assistment \supseteq Problem$

Từ cấu trúc này, dễ dàng nhận ra rằng mỗi công việc có một hoặc nhiều chỉ dẫn, mỗi chỉ dẫn có nhiều vấn đề. Nhiệm vụ là dự đoán đúng năng lực (kết quả) của sinh viên khi lần đầu sinh viên giải quyết vấn đề. Hình 4.6 là một mẫu dữ liệu ví dụ được sử dụng thực nghiệm. Tập dữ liệu ASSISTments sau khi xử lý bao gồm 8519 sinh viên (8519 users), 35978 bài tập (35978 items), 1011079 đánh giá năng lực (1011079 ratings).

	A	B	C	D	E	F	G
1	user_id	student_class_id	assignment_id	assistment_id	problem_id	skill_id	correct
2	77759	12138	245748	12914	12914	231	1
3	77759	12138	245748	15320	15320	231	1
4	77759	12138	245748	14529	14529	231	1
5	77912	12138	245698	1159	1159	100	0
6	77912	12138	245698	1647	1647	93	1
7	77912	12138	245698	2705	2705	100	1
8	77912	12138	245698	2186	2186	35	1
9	77912	12138	245698	1653	1653	35	1
10	77912	12138	245698	2345	2345	35	1
11	77912	12138	245698	2196	2196	35	1
12	77912	12138	245698	3522	3522	101	1

Hình 4.6: Mẫu dữ liệu của tập dữ liệu ASSISTments

4.4.2 Đánh giá mô hình

Để đánh giá hiệu quả của giải thuật, chúng tôi sử dụng nghi thức kiểm tra hold-out: lấy ngẫu nhiên 2/3 tập dữ liệu để học và 1/3 còn lại để 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), nên có hai cách đánh giá phù hợp nhất là: Root Mean Squared Error (RMSE) và Mean Absolute Error (MAE). Trong bài toán này, dự đoán điểm sinh viên là nhiệm vụ dự đoán xếp hạng (phản hồi rõ ràng), vì vậy thước đo phổ biến trong RS là Root Mean Squared Error (RMSE) được sử dụng để đánh giá mô hình.

4.4.3 Tham số thực nghiệm

Độ chính xác của dự đoán phụ thuộc rất nhiều vào các tham số cung cấp cho thuật toán. Vì vậy, việc tìm kiếm các tham số tốt nhất là rất quan trọng. Phương pháp tìm kiếm siêu tham số được áp dụng để tìm kiếm tất cả các tham số của các mô hình được tiếp cận. Tìm kiếm siêu tham số có hai giai đoạn dựa trên tìm kiếm lưới: tìm kiếm thô (đối với các đoạn dài) và tìm kiếm mịn (đối với các đoạn ngắn). Đầu tiên, một tìm kiếm thô được áp dụng để tìm các siêu tham số tốt nhất trong các phân đoạn dài. Sau đó, chúng tôi sử dụng một tìm kiếm mịn để tìm các siêu tham số gần đó. Ví dụ: sử dụng RMSE làm tiêu chí, kết quả tìm kiếm siêu tham số cho các mô hình trên tập dữ liệu ASSISTments được trình bày trong Bảng 4.1.

Bảng 4.1: Các tham số của giải thuật tích hợp kiến trúc học sâu thực nghiệm trên tập dữ liệu ASSISTments

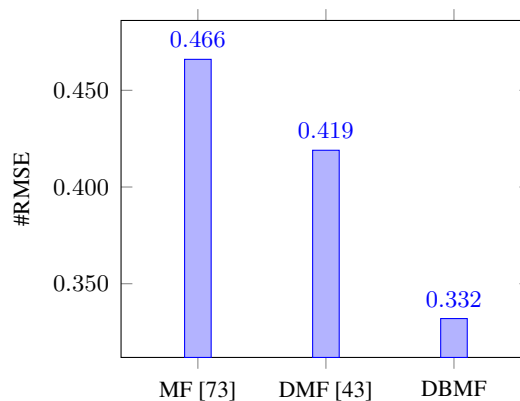
Phương pháp	Các siêu tham số	Thời gian	Độ phức tạp
MF	$\beta = 0.03, \#i = 50, k = 4, \lambda = 0.05$	100s	$O(s \times c \times i)$
DMF	$Deep(d) = 4$ $numFactors(k) = [3, 6, 3, 3];$ $numIters(i) = [50, 50, 50, 50];$ $learningRate(\beta) = [0.01, 0.1, 0.1, 0.1];$ $reg(\lambda) = [0.01, 0.1, 0.1, 0.01]$	24h	$O(s \times c \times i \times d)$
DBMF	$Deep(d) = 4$ $numFactors(k) = [3, 6, 3, 3];$ $numIters(i) = [50, 50, 50, 50];$ $learningRate(\beta) = [0.01, 0.1, 0.1, 0.1];$ $reg(\lambda) = [0.01, 0.1, 0.1, 0.01]$	36h	$O(s \times c \times i \times d)$

Chúng tôi sử dụng các siêu tham số tốt nhất tìm được để đào tạo và thử nghiệm từng mô hình tương ứng. Tuy nhiên, thời gian đào tạo chậm hơn so với không học sâu do thời gian huấn luyện qua nhiều tầng và thời gian tìm kiếm siêu tham số tối ưu rất lâu.

4.4.4 Kết quả thực nghiệm

Nghiên cứu đã so sánh việc sử dụng kiến trúc học sâu tích hợp vào phân rã ma trận (DMF), cũng như phân rã sâu ma trận thiên vị (DBMF) để giải quyết vấn đề dự đoán kết quả học tập của sinh viên với các phương pháp MF và Biased-MF.

Các thí nghiệm được tiến hành với bốn phương pháp, kết quả thử nghiệm được hiển thị trong (Hình 4.7). So sánh với các kết quả khác, RMSE của phương pháp tiếp cận được đề xuất (thuật toán 4.2) là nhỏ nhất (0,332) trên tập dữ liệu. Sai số nhỏ nhất chứng tỏ mô hình tốt nhất.



Hình 4.7: Kết quả thực nghiệm trên tập dữ liệu ASSISTment

Các phương pháp để so sánh dùng trong thực nghiệm

- **MF:** Phương pháp lọc cộng tác dựa vào mô hình phân rã ma trận [73].
- **DMF:** Phương pháp lọc cộng tác dựa vào mô hình phân rã sâu ma trận [43].
- **DBMF:** Phương pháp lọc cộng tác dựa vào mô hình phân rã sâu ma trận thiên vị.

4.5 Tổng kết chương

Chương này đã giới thiệu một cách tiếp cận sử dụng kiến trúc học sâu để tích hợp phương pháp MF nhằm nâng cao độ chính xác của dự đoán kết quả học tập của sinh viên. Chúng ta có thể tận dụng các nguyên tắc học sâu để tinh chỉnh đầu ra của mô hình

(ma trận dự đoán) thông qua đào tạo liên tiếp để xây dựng mô hình dự đoán với cách tiếp cận này. Do đó, kết quả dự đoán có thể được cải thiện đáng kể. Tiến hành thử nghiệm trên tập dữ liệu chuẩn cho thấy rằng quy trình đề xuất hoạt động tốt. Không dừng lại ở đó, chúng tôi còn đề xuất cải tiến mô hình phân rã sâu ma trận với tham số đầu vào là một ma trận thiên vị, gọi là giải thuật DBMF. Kết quả thực nghiệm cho thấy một lần nữa việc ứng dụng kiến trúc học sâu hay phân rã sâu, phân rã nhiều lần một ma trận sẽ tăng độ chính xác dự đoán cho mô hình.

Việc áp dụng kiến trúc học sâu để phân rã ma trận khiến thời gian đào tạo sẽ bị chậm lại (thời gian thực hiện dự đoán có thể là một hai ngày). Tuy nhiên, ta dễ dàng thực hiện một thuật toán song song để giải quyết vấn đề này. Thêm vào đó, đây là bài toán dự đoán kết quả học tập của sinh viên, việc dự đoán này chỉ thực hiện 1 lần trong cả học kỳ (từ 3 đến 6 tháng) rồi mới đưa kết quả dự đoán vào hệ thống gợi ý cho sinh viên lựa chọn môn học tự chọn trong lập kế hoạch học tập. Do vậy thời gian thực hiện thuật toán không ảnh hưởng nhiều trong trường hợp này, quan trọng là độ chính xác dự đoán càng chính xác là càng tốt.

Tuy phương pháp ứng dụng ý tưởng của kiến trúc học sâu vào kỹ thuật phân rã ma trận đạt được kết quả khả quan, nhưng phương pháp này chưa tận dụng được các thông tin bổ sung (meta-data) khác nhằm nâng cao hiệu quả của dự đoán. Chính vì thế, luận án sẽ tiếp tục trình bày nghiên cứu phương pháp tích hợp các mối quan hệ dữ liệu vào dự đoán kết quả học tập của sinh viên ở Chương 5 tiếp theo.

CHƯƠNG 5 TÍCH HỢP CÁC MỐI QUAN HỆ DỮ LIỆU VÀO DỰ ĐOÁN KẾT QUẢ HỌC TẬP CỦA SINH VIÊN

Chương này sẽ trình bày chi tiết các phương pháp nâng cao độ chính xác của giải thuật phân rã ma trận như: tích hợp mối quan hệ xã hội của sinh viên (dựa trên mối quan hệ bạn bè cùng lớp học) và mối liên quan giữa các môn học (dựa trên kiến thức, kỹ năng được miêu tả trên các thuộc tính của môn học). Hai phương pháp tích hợp được thiết kế trực tiếp lên giải thuật cũng như công thức tính toán nhằm cải thiện hiệu quả dự đoán, chúng bổ sung dữ liệu cho kỹ thuật phân rã ma trận từ nguồn khác, trong khi các giải thuật khác của lọc cộng tác như người dùng tương tự, đối tượng tự lại sử dụng lại các thông tin đánh giá hiện có của ma trận đánh giá. Kết quả được nghiệm trên các bộ dữ liệu công bố và nội bộ, đều cho kết quả độ tốt hơn, điều này có thể cho thấy thuật toán khả thi trong việc ứng dụng vào bài toán dự đoán kết quả học tập của sinh viên. Nghiên cứu này được công bố trong các công trình [CT2][CT3].

5.1 Đặt vấn đề

Tuy phương pháp ứng dụng ý tưởng của kiến trúc học sâu vào kỹ thuật phân rã ma trận (được trình bày ở chương 4) đạt được kết quả khả quan, nhưng phương pháp này chưa tận dụng được các thông tin bổ sung khác nhằm nâng cao hiệu quả của dự đoán. Chính vì thế, trong chương này, luận án sẽ tiếp tục trình bày nghiên cứu phương pháp tích hợp các mối quan hệ dữ liệu vào dự đoán kết quả học tập của sinh viên

Lợi thế về mặt dữ liệu phù hợp với việc tích hợp các mối quan hệ dữ liệu vào mô hình dự đoán năng lực học tập của sinh viên bao gồm hai khía cạnh: Một, dữ liệu mạng xã hội của người dùng - user (hay đơn giản hơn chỉ là ma trận quan hệ bạn bè) để thể hiện mối quan hệ người dùng ảnh hưởng đến kết quả dự đoán với giả thuyết rằng nếu sinh viên có mối quan hệ bạn bè với nhau sẽ có kết quả học tập ảnh hưởng lẫn nhau; Hai, mối liên quan giữa các mục tin -item (hay mối liên quan giữa kiến thức yêu cầu, chuẩn đầu ra của môn học) để thể hiện mối liên quan giữa các môn học nhằm nâng cao hiệu quả dự đoán

với giả thuyết tương tự.

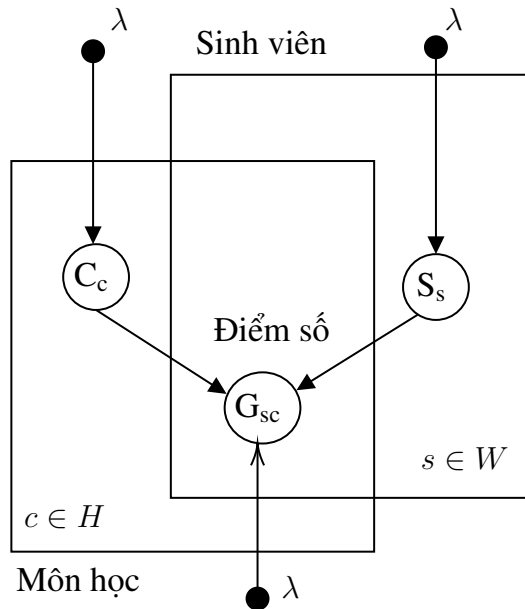
Trong tập dữ liệu ASSISTment có trường `student_class_id` (lớp học của sinh viên) để tìm mối quan hệ bạn bè của sinh viên và trường `skill_id` (kỹ năng cần đạt của sinh viên) để tìm mối liên quan giữa các môn học. Hai trường này rất có ý nghĩa trong việc nâng cao hiệu quả dự đoán.

5.2 Phương pháp tích hợp mối quan hệ người dùng

5.2.1 Phương pháp phân rã ma trận chưa tích hợp mối quan hệ

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 [73] được minh họa trong Hình 5.1, nghĩa là $R \approx WH^T$.

$W \in R^{|S| \times |K|}$ 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 $H \in R^{|C| \times |K|}$ 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 . Mô hình mô tả giải thuật phân rã ma trận cơ sở được trình bày ở Hình 5.1.



Hình 5.1: Mô hình đồ họa của kỹ thuật phân rã ma trận cơ sở

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 (5.1) như sau:

$$\hat{g}_{sc} = \sum_{k=1}^K w_{sk} h_{ck} = w_s \bullet h_c^T \quad (5.1)$$

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 theo điều kiện nào đó, chẳng hạn RMSE (Root Mean Squared Error) theo công thức (5.2) sau:

$$RMSE = \sqrt{\frac{1}{|D^{\text{test}}|} \sum_{s,c,g \in D^{\text{test}}} (g_{sc} - \hat{g}_{sc})^2} \quad (5.2)$$

Tối ưu hóa hàm mục tiêu theo phương pháp giảm dốc ngẫu nhiên (Stochastic Gradient Descent - SGD) [3] như công thức (5.3) sau:

$$o^{MF} = \sum_{(s,c,g) \in D^{\text{tan}}} \left(g_{sc} - \sum_{k=1}^K w_{sk} h_{ck} \right)^2 + \lambda (\|W\|_F^2 + \|H\|_F^2) \quad (5.3)$$

Với λ là hệ số chính tắc hóa ($0 \leq \lambda < 1$) và $\|\cdot\|_F^2$ là chuẩn Frobenius. Đại lượng $\lambda (\|W\|_F^2 + \|H\|_F^2)$ được 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 (5.4) và (5.5) sau:

$$w'_{sk} = w_{sk} + \beta (2e_{sc} h_{ck} - \lambda w_{sk}) \quad (5.4)$$

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

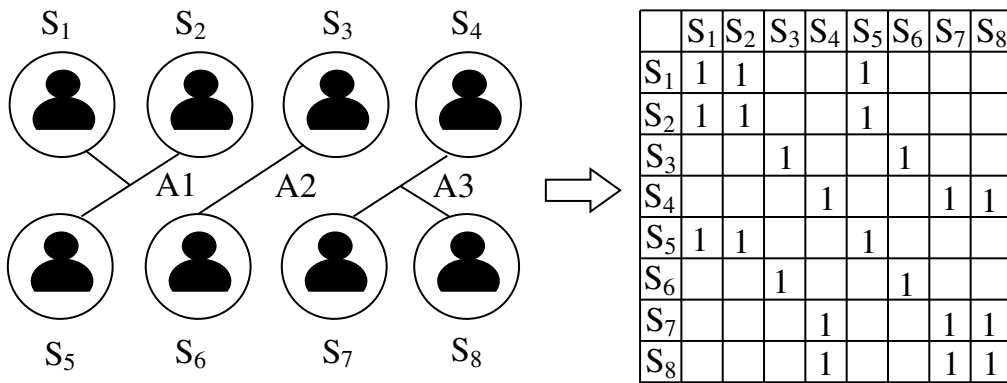
Trong đó: $e_{sc} = g_{sc} - \hat{g}_{sc}$ là độ lỗi của dự đoán, β là tốc độ học. Sau quá trình huấn luyện ta được hai ma trận W và H đã tối ưu, thì quá trình dự đoán được thực hiện. Quá trình dự đoán được tính và biểu diễn như công thức (5.6) sau:

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

5.2.2 Phương pháp tích hợp mối quan hệ bạn bè

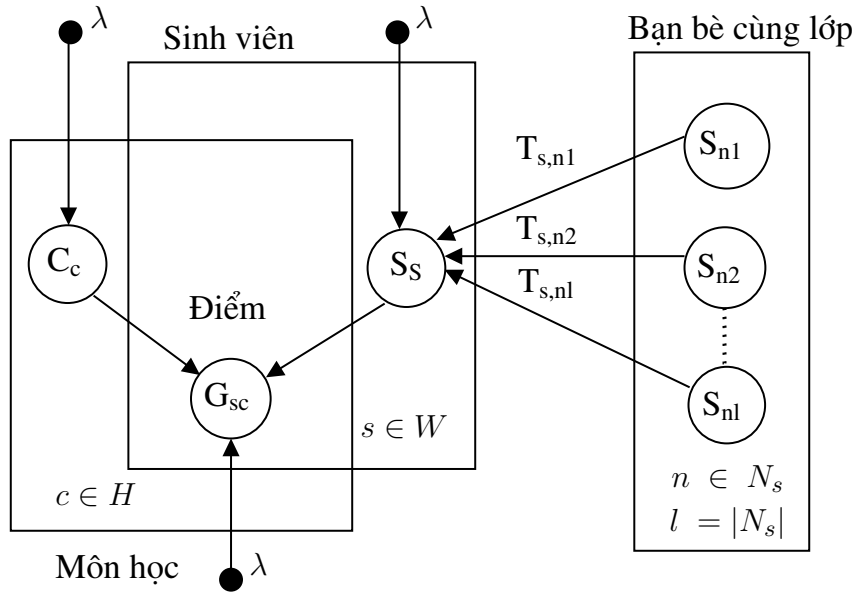
Với giả định rằng “Nếu A là bạn của B, và A học tốt cũng như chăm học thì có thể ảnh hưởng đến B và ngược lại”. Để thực nghiệm giả định này, nghiên cứu [77] đã tích hợp mối quan hệ bạn bè của sinh viên vào mô hình dự đoán năng lực học tập của sinh viên, và việc tích hợp này đã cho kết quả tốt hơn.

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 (Hình 5.2). Trong ví dụ này, có 3 sinh viên (S_1, S_2, S_5) là bạn của nhau do học lớp A1, tương tự như vậy (S_3, S_6) học chung lớp A2, và (S_4, S_7, S_8) học lớp A3. Ta chỉ xét là có hoặc không có mối quan hệ bạn bè trong một lớp học để chuyển chúng thành ma trận mối quan hệ, mà chưa quan tâm đến giá trị trọng số ảnh hưởng của mối quan hệ bạn bè này.



Hình 5.2: Sự chuyển đổi mối quan hệ bạn bè thành ma trận mối quan hệ

Trong hướng tiếp cận tích hợp quan hệ người dùng, 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 0 hoặc 1. Nếu $T = 0$ thì hai người s và n khác lớp, nếu $T = 1$ thì hai người học chung lớp. Hình 5.3 thể hiện mô hình tích hợp mối quan hệ người dùng vào kỹ thuật MF.



Hình 5.3: Mô hình tích hợp mối quan hệ người dùng vào kỹ thuật phân rã ma trận

Mối quan hệ của một sinh viên được biểu diễn bằng một véc-tơ chứa các giá trị quan hệ của n sinh viên chung lớp với sinh viên s , tập mối quan hệ này được biểu diễn bằng ma trận $T^{|S||S|}$, là một ma trận nhị phân đối xứng.

Để tích hợp được mối quan hệ trong lớp học của sinh viên vào MF, chúng ta chỉ cần thay thế ma trận W bằng $\hat{W}$ bằng công thức (5.7) và sau đó thực hiện tương tự kỹ thuật MF cơ sở.

$$\hat{w}_s = \frac{\sum_{n \in N_s} T_{s,n} w_n}{\sum_{n \in N_s} T_{s,n}} = \frac{1}{|N_s|} \sum_{n \in N_s} w_n \quad (5.7)$$

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 công thức của giải thuật MF nhưng thay thế W bằng $\hat{W}$ Như vậy, công thức (5.8) tính hàm dự đoán trở thành:

$$\hat{g}_{sc} = \sum_{k=1}^K \hat{w}_{sk} h_{ck} = \hat{w}_s \bullet h_c^T \quad (5.8)$$

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

được viết theo công thức (5.9) như sau:

$$O^{\text{SocialMF}} = \sum_{(s,c,g) \in D^{\text{train}}} \left(g_{sc} - \sum_{k=1}^K \hat{w}_{sk} h_{ck} \right)^2 + \lambda (\|W\|_F^2 + \|H\|_F^2) + \lambda_T \sum_{s=1}^S \left(w_s - \frac{1}{|N_s|} \sum_{n \in N_s} w_n \right)^2 \quad (5.9)$$

Như vậy, các giá trị w và h cũng được cập nhật mới sau mỗi bước lặp được thực hiện theo 2 công thức (5.10) và (5.11) như sau:

$$w'_{sk} = w_{sk} + \beta (2e_{sc} h_{ck} - \lambda w_{sk}) + \lambda_T \left(w_{sk} - \frac{1}{|N_s|} \sum_{n \in N_s} w_{nk} \right) - \lambda_T \left(\sum_{t \in S_s} \frac{1}{|N_s|} \left(w_{tk} - \frac{1}{|N_s|} \sum_{w \in N_s} w_{wk} \right) \right) \quad (5.10)$$

$$h'_{ck} = h_{ck} + \beta (2e_{sc} w_{sk} - \lambda h_{ck}) \quad (5.11)$$

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 mỗi 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 được các giá trị của ma trận W và H đã tối ưu. Khi đó, chúng ta có thể dễ dàng dự đoán kết quả thông qua công thức (5.12):

$$\hat{g}_{sc} = \sum_{k=1}^K w_{sk} h_{ck} = w_s \bullet h_c^T \quad (5.12)$$

Chi tiết cho kỹ thuật tích hợp mỗi quan hệ bạn bè được tóm tắt trong (thuật toán 5.1) (Predicting Student Performance - SocialMF). Trước tiên, ta cần khởi tạo giá trị của các tham số một cách ngẫu nhiên theo chuẩn phân phối $N(\mu, \sigma^2)$ với giá trị trung bình $\mu = 0$, độ lệch chuẩn $\sigma^2 = 0.01$. Khi điều kiện dừng chưa thỏa mãn, chẳng hạn như đạt đến số lần lặp tối đa hoặc tối điểm hội tụ thì các tham số vẫn còn cập nhật (converging: $O^{\text{SocialMF}}_{\text{Iter}_{n-1}} - O^{\text{SocialMF}}_{\text{Iter}_n} < \varepsilon$). Các dòng lệnh từ 12 đến dòng 22 là cập nhật cho nhân tố tiềm ẩn của sinh viên theo công thức (5.10), và dòng lệnh số 23 cập nhật cho nhân tố tiềm ẩn của môn học theo công thức (5.11). Sau khi cập nhật các ma trận nhân tố w và h cho đến điều kiện được thỏa mãn, thủ tục được trả về hai ma trận sau khi đã tối ưu và sẵn sàng cho dự đoán ở bước tiếp theo.

Giải thuật 5.1 Predicting Student Performance - SocialMF

Data: D^{train} , K , betas, lamdas, stopping condition, T-UserRelation

Result: W, H

```

1 initialization
2 Let  $s \in S$  be a student,  $c \in C$  be a course,  $g \in G$  be a grade
3 Let  $W[|S|][K], H[|C|][K]$  be latent factors of students, courses
4 Let  $N_s$  is the number of students who are friends with  $s \in T\text{-UserRelation}$ 
5  $W \leftarrow N(0, \sigma^2)$  and  $H \leftarrow N(0, \sigma^2)$ 
6  $k, t, \beta, \lambda \leftarrow \text{pop}(K, T, \text{Betas}, \text{Lamdas})$ 
7 while Stopping criterion is NOT met do
8   foreach  $(s, c, g_{sc})$  in  $D^{train}$  do
9      $\hat{g}_{sc} \leftarrow \sum_k^K (W[s][k] * H[c][k])$ 
10     $e_{sc} = g_{sc} - \hat{g}_{sc}$ 
11    for  $k \leftarrow 1$  to  $K$  do
12      for  $n \leftarrow 1$  to  $N_s$  do
13         $VT = VT - W[n][k]/N_s$ 
14      end
15       $VT = W[s][k] - VT$ 
16      for  $t \leftarrow 1$  to  $S$  do
17        for  $w \leftarrow 1$  to  $N_s$  do
18           $W[t][k] = W[t][k] - W[w][k]/N_s$ 
19        end
20         $VP = VP + W[t][k]/N_s$ 
21      end
22       $W[s][k] \leftarrow W[s][k] + \beta * (2e_{sc} * H[c][k] - \lambda * W[s][k]) + \lambda_T(VT - VP)$ 
23       $H[c][k] \leftarrow H[c][k] + \beta * (2e_{sc} * W[s][k] - \lambda * H[c][k])$ 
24    end
25  end
26 end
27 return ( $< W, H >$ )

```

Độ phức tạp thời gian của thuật toán 5.1 này phụ thuộc chủ yếu vào số lần lặp, kích thước của ma trận đánh giá và số lần tính toán mối quan hệ của sinh viên. Với số lượng sinh viên là s , số lượng môn học là c , số lượng lần lặp là i , số người bạn của sinh viên là $s - 1$ và số lượng nhân tố tiềm ẩn là k (do k là hằng số rất nhỏ nên có thể bỏ qua), và tương tự như vậy theo nguyên tắc lấy tối đa. Do đó độ phức tạp thuật toán là $O(s \times c \times i \times s \times (s - 1)) \approx O(s^3 \times c \times i)$

5.2.3 Đánh giá kết quả thực nghiệm phương pháp tích hợp mối quan hệ người dùng

Không giống như giải thuật student-kNNs sẽ tìm độ tương đồng của sinh viên được đánh giá năng lực qua điểm số. Trong khi đó, giải thuật SocialMF đã tận dụng được mối quan hệ bạn bè trong lớp (bổ sung từ nguồn khác) nhằm tăng độ chính xác cho dự đoán.

5.2.3.1 Tập dữ liệu thực nghiệm

Chúng tôi sử dụng hai tập dữ liệu để thực nghiệm là: ASSISTments và CTU.

Tập dữ liệu ASSISTments sau khi xử lý bao gồm 8519 sinh viên (8519 users), 35978 bài tập (35978 items), 1011079 đánh giá năng lực (1011079 ratings).

Tập dữ liệu điểm của khoa CNTT&TT, trường Đại học Cần Thơ (CTU) sau khi xử lý bao gồm 4017 sinh viên (4017 user) và 353 môn học (353 item) của 3 ngành học và gồm 279536 điểm chi tiết (279536 ratings).

5.2.3.2 Nghi thức kiểm tra và độ đo

Để đánh giá hiệu quả của giải thuật, chúng tôi 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.

Có rất nhiều phương pháp dùng để đánh giá hiệu quả của giải thuật gợi ý nhưng phải phụ thuộc vào dạng bài toán. Do đó, khi thực hiện đánh giá giải thuật cần chọn phương pháp đánh giá phù hợp với giải thuật và cả dữ liệu bài toán. Tuy nhiên do bài toán dự đoán kết quả học tập của sinh viên thuộc dạng rating prediction (dự đoán từ đánh giá tường minh), nên có hai cách đánh giá phù hợp nhất là: Root Mean Squared Error (RMSE) và Mean Absolute Error (MAE) được biểu diễn ở hai công thức (5.13) và (5.14):

$$RMSE = \sqrt{\frac{1}{|D^{test}|} \sum_{s,c,g \in D^{test}} (g_{sc} - \hat{g}_{sc})^2} \quad (5.13)$$

$$MAE = \frac{1}{|D^{test}|} \sum_{s,c,g \in D^{test}} |(g_{sc} - \hat{g}_{sc})| \quad (5.14)$$

Hơn nữa, các giải thưởng lớn trong lĩnh vực RS đều dùng RMSE để đánh giá, như Netflix Prize, KDD Cup 2010,...

5.2.3.3 Các tham số trong thực nghiệm

Các siêu tham số (hyper-parameters) của MF, BMF, SocialMF như số lần lặp (Iter), số nhân tố tiềm ẩn k , tốc độ học β , và hệ số chính tắc hóa λ được xác định bằng phương pháp tìm kiếm siêu tham số. Các tham số sử dụng trong thực nghiệm được mô tả chi tiết trong hai bảng 5.1 và bảng 5.2.

Bảng 5.1: Tham số thực nghiệm trên tập dữ liệu CTU

Phương pháp	Giá trị các siêu tham số
Student-kNNs	$K = 5, simMeasure = Cosine$
MF	$\beta = 0.01, \#iter = 30, K = 10, \lambda = 0.049$
Social-MF	$K = 80, \lambda = 3, \beta = 0.01, SocialReg = 5, Iter = 500$

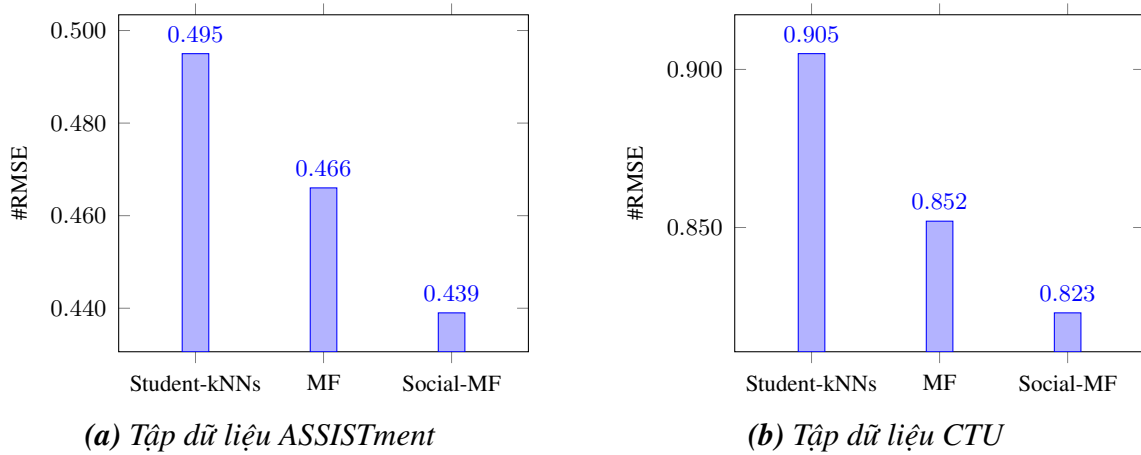
Bảng 5.2: Tham số thực nghiệm trên tập dữ liệu ASSISTment

Phương pháp	Giá trị các siêu tham số
Student-kNNs	$K = 5, simMeasure = Cosine$
MF	$\beta = 0.01, \#iter = 60, K = 32, \lambda = 0.1$
Social-MF	$K = 80, \lambda = 3, \beta = 0.01, SocialReg = 5, Iter = 500$

5.2.3.4 Kết quả thực hiện

Từ biểu đồ so sánh (Hình 5.4) cho thấy khi áp dụng giải thuật tích hợp mối quan hệ của người dùng vào bài toán dự đoán kết quả học tập của sinh viên được thực nghiệm

trên hai tập dữ liệu ở Việt Nam và Quốc tế đều đạt được độ lỗi RMSE thấp nhất so với các giải thuật khác.



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

Qua những thí nghiệm cho thấy giải thuật SocialMF mang lại hiệu quả cao hơn các phương pháp chưa tích hợp (MF) và phương pháp người dùng tương đồng (Student-kNNs) trên cả hai tập dữ liệu Việt Nam và Quốc tế. Thời gian thực hiện việc tích hợp không đáng kể, tuy nhiên việc xử lý dữ liệu mỗi quan hệ phụ thuộc và số lượng người dùng (sinh viên) và được thực hiện riêng lẻ với việc huấn luyện ma trận đánh giá.

5.3 Phương pháp tích hợp mối liên quan các đối tượng

5.3.1 Phương pháp tích hợp mối liên quan các môn học

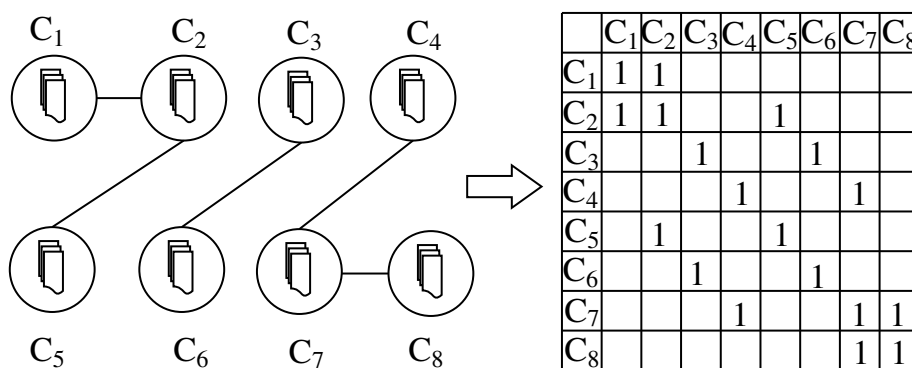
Theo đào tạo tín chỉ, chương trình đào tạo thường xuyên thay đổi, có thể phát sinh những môn học mới, nhưng nội dung của môn học này có thể liên quan đến các môn học khác. Như vậy, các môn học có mối quan hệ mật thiết với nhau. Với ý tưởng tương tự: “Nếu hai vấn đề A và B có yêu cầu kỹ năng giống nhau, 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.

Có hai cách để tìm mối tương quan giữa các môn học: (1) tìm thấy ở phân loại từ người quản lý đào tạo (nhóm môn học, môn học trước, môn học tiên quyết); (2) tìm thấy từ việc xét nội dung tương đương nhau nhờ vào mô tả của môn học trong đề cương chi tiết. Tùy vào cấu trúc của tập dữ liệu ta có thể sử dụng những phương pháp phù hợp.

Đối với tập dữ liệu ASSISTments thì trong tập dữ liệu không có mô tả nội dung môn học nhưng có trường Skill_id chỉ kỹ năng cần đạt trong câu hỏi đó và chúng tôi sẽ sử dụng chúng như kiểu phân loại đơn giản, những câu hỏi có cùng kỹ năng sẽ xem như liên quan nhau.

Một số nghiên cứu khác đã tìm kiếm mối liên quan giữa các mục tin (item) để tích hợp vào MF nhằm nâng cao hiệu quả cho giải thuật [74] [78]. Tác giả đã sử dụng lại việc tìm những mục tin có đánh giá gần giống nhau (similar item) để nâng cao hiệu quả dự đoán, với nghiên cứu này phải phụ thuộc vào việc người dùng có đánh giá các mục tin hay không. Nghiên cứu [48] đã kết hợp các thuộc tính vào mục tin để áp dụng kỹ thuật MF, cải tiến này phù hợp cho dữ liệu mục tin có nhiều thông tin và bỏ qua sự liên quan giữa các mục tin. Tuy nhiên đây cũng là những nghiên cứu thành công trong việc tìm kiếm khai thác thông tin bổ sung từ mục tin để nâng cao độ chính xác của dự đoán.

Trước khi tích hợp mối quan hệ giữa các môn học vào hệ thống dự đoán kết quả học tập của sinh viên [79], ta cần chuyển mối quan hệ này về ma trận quan hệ nhị phân R . Với hướng tích hợp này, mỗi môn học có tập hợp R_c thể hiện mối quan hệ 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. Hình 5.5 thể hiện sự chuyển đổi từ mối quan hệ sang ma trận nhị phân. Cũng tương tự như mối quan hệ người dùng. Trong ví dụ này, ta có ba môn học: môn thứ nhất C_1, C_2, C_5 , môn thứ hai, C_3 và C_6 , môn thứ ba C_4, C_7, C_8 là có liên quan kiến thức với nhau, tương tự như vậy ta thực hiện hết tập dữ liệu về môn học để xây dựng ma trận mối liên quan giữa các môn học này để tích hợp vào mô hình.



Hình 5.5: Mô hình chuyển đổi quan hệ môn học thành ma trận

Mô hình phân rã ma trận trong hệ thống gợi ý được sử dụng làm nền tảng cho việc tích hợp mối quan hệ giữa các môn học, việc tích hợp mối quan hệ được gắn vào mô

Như vậy, các giá trị w và h cũng được cập nhật mới sau mỗi bước lặp được thực hiện

theo biểu thức (5.18) và (5.19) như sau, trong đó $e_{sc} = g_{sc} - \hat{g}_{sc}$

$$w'_{sk} = w_{sk} + \beta (2e_{sc}h_{ck} - \lambda w_{sk}) \quad (5.18)$$

$$\begin{aligned} h'_{ck} = & h_{ck} + \beta (2e_{sc}w_{sk} - \lambda h_{ck}) \\ & + \lambda_R \left(h_{ck} - \frac{1}{|M_c|} \sum_{m \in M_c} h_{mk} \right) \\ & - \lambda_R \left(\sum_{r \in C_c} \frac{1}{|M_c|} \left(h_{rk} - \frac{1}{|M_c| \sum_{y \in R_c} h_{yk}} \right) \right) \end{aligned} \quad (5.19)$$

Với β là tốc độ học, λ là hệ số chính tắc cho ma trận W và H , λ_R là hệ số chính tắc cho ma trận mối 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 (5.20) dự đoán kết quả được trình bày:

$$\hat{g}_{sc} = w_s \cdot \hat{h}_c^T \quad (5.20)$$

Tương tự ý tưởng của việc tích hợp mối qua hệ bạn bè (thuật toán 5.1), ta mở rộng và thực hiện tương tự cho việc tích hợp mối liên quan giữa các môn học được gọi là thuật toán PredictingStudentPerformance-CRMF (thuật toán (5.2)) sẽ thiết kế cho cho việc tích hợp mối liên quan giữa các môn học, được trình bày chi tiết như sau: Trước tiên, ta cũng khởi tạo giá trị của các tham số một cách ngẫu nhiên theo chuẩn phân phối $N(\mu, \sigma^2)$ với giá trị trung bình $\mu = 0$, độ lệch chuẩn $\sigma^2 = 0.01$. Khi điều kiện dừng chưa thỏa mãn, chẳng hạn như đạt đến số lần lặp tối đa hoặc tới điểm hội tụ thì các tham số vẫn còn cập nhật (converging: $O_{Iter_{n-1}}^{CRMF} - O_{Iter_n}^{CRMF} < \varepsilon$). Các dòng lệnh từ 12-21 và dòng 23 là cập nhật cho nhân tố tiềm ẩn của môn học theo công thức (5.19), và dòng lệnh số 22 cập nhật cho nhân tố tiềm ẩn của sinh viên theo công thức (5.18). Sau khi cập nhật các ma trận nhân tố w và h cho đến điều kiện được thỏa mãn, thủ tục được trả về hai ma trận sau khi đã tối ưu và sẵn sàng cho dự đoán ở bước tiếp theo.

Giải thuật 5.2 PredictingStudentPerformance-CRMF

Data: D^{train} , K , β , λ , lamdas , stopping condition, R -CoursesRelation

Result: W, H

```

1 initialization
2 Let  $s \in S$  be a student,  $c \in C$  be a course,  $g \in G$  be a grade
3 Let  $W[|S|][K], H[|C|][K]$  be latent factors of students, courses
4 Let  $R_c$  is the number of courses which are related with  $c \in R$ -CoursesRelation
5  $W \leftarrow N(0, \sigma^2)$  and  $H \leftarrow N(0, \sigma^2)$ 
6  $k, t, \beta, \lambda \leftarrow \text{pop}(K, T, \text{Betas}, \text{Lamdas})$ 
7 while Stopping criterion is NOT met do
8   foreach  $(s, c, g_{sc})$  in  $D^{train}$  do
9      $\hat{g}_{sc} \leftarrow \sum_k^K (W[s][k] * H[c][k])$ 
10     $e_{sc} = g_{sc} - \hat{g}_{sc}$ 
11    for  $k \leftarrow 1$  to  $K$  do
12      for  $m \leftarrow 1$  to  $R_c$  do
13         $VT = VT - H[m][k]/R_c$ 
14      end
15       $VT = H[c][k] - VT$ 
16      for  $r \leftarrow 1$  to  $C$  do
17        for  $y \leftarrow 1$  to  $R_c$  do
18           $H[r][k] = H[r][k] - H[y][k]/R_c$ 
19        end
20         $VP = VP + H[t][k]/R_c$ 
21      end
22       $W[s][k] \leftarrow W[s][k] + \beta * (2e_{sc} * H[c][k] - \lambda * W[s][k])$ 
23       $H[c][k] \leftarrow H[c][k] + \beta * (2e_{sc} * W[s][k] - \lambda * H[c][k]) + \lambda_R(VT - VP)$ 
24    end
25  end
26 end
27 return  $(W, H)$ 

```

Độ phức tạp thời gian của thuật toán 5.2 này phụ thuộc chủ yếu vào số lần lặp, kích thước của ma trận đánh giá và số lần tính toán mối quan hệ của sinh viên. Với số lượng sinh viên là s , số lượng môn học là c , số lượng lần lặp là i , số người bạn của sinh viên là $s - 1$ và số lượng nhân tố tiềm ẩn là k (do k là hằng số rất nhỏ nên có thể bỏ qua), và tương tự như vậy theo nguyên tắc lấy tối đa. Do đó độ phức tạp thuật toán là $O(s \times c \times i \times c \times (c - 1)) \approx O(c^3 \times s \times i)$

5.3.2 Đánh giá kết quả thực nghiệm

5.3.2.1 Tập dữ liệu thực nghiệm

Tập dữ liệu ASSISTments sau khi xử lý gồm 8519 sinh viên (8519 users), 35978 bài tập (35978 items), 1011079 đánh giá năng lực (1011079 ratings).

5.3.2.2 Nghi thức kiểm tra và độ đo

Để đánh giá hiệu quả của giải thuật, chúng tôi 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), nên chúng tôi sử dụng độ đo Root Mean Squared Error (RMSE) để thực nghiệm được biểu diễn ở công thức (5.21):

$$RMSE = \sqrt{\frac{1}{|D^{\text{test}}|} \sum_{s,c,g \in D^{\text{test}}} (g_{sc} - \hat{g}_{sc})^2} \quad (5.21)$$

5.3.2.3 Tìm kiếm siêu tham số

Tìm kiếm siêu tham số (hyper-parameter search) [74] để giải thuật đạt hiệu quả cao. Chúng ta thực hiện việc tìm kiếm này qua hai giai đoạn: tìm thô (Coarse Search) và tìm mịn (Granularity Search).

Các siêu tham số (hyper-parameters) của MF, SocialMF, CRMF như số lần lặp (Iter), số nhân tố tiềm ẩn k , tốc độ học β , và hệ số chính tắc hóa λ được xác định bằng phương pháp tìm kiếm siêu tham số. Các tham số sử dụng trong thực nghiệm được mô tả chi tiết trong bảng 5.3.

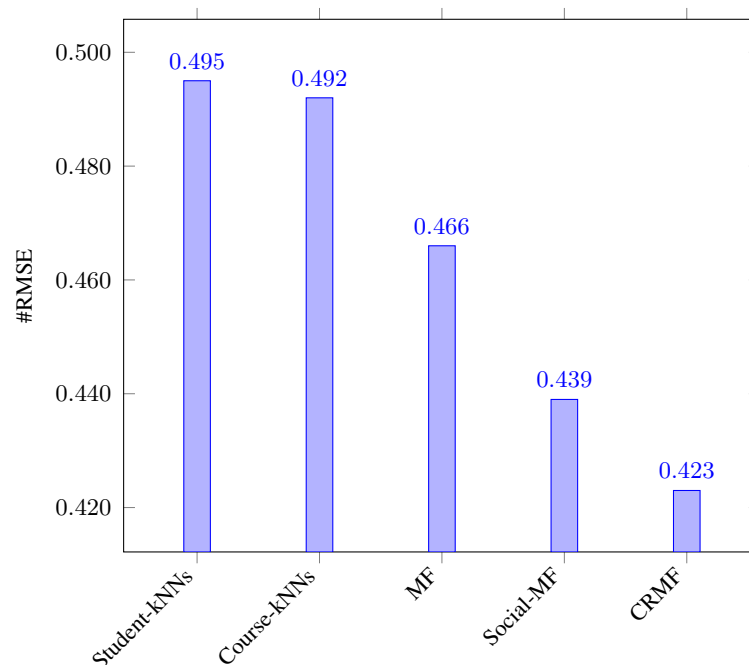
Bảng 5.3: Các tham số thực nghiệm của các giải thuật trên tập ASSISTments

Phương pháp	Giá trị các siêu tham số	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	$\beta = 0.01, \#i = 60, k = 32, \lambda = 0.1$	$O(s \times c \times i)$
Social-MF	$k = 80, \lambda = 3, \beta = 0.01, SocialReg = 5, i = 500$	$O(s^3 \times c \times i)$
CRMF	$k = 80, \lambda = 3, \beta = 0.01, ItemReg = 3, i = 300$	$O(c^3 \times s \times i)$

5.3.2.4 Kết quả thực hiện

Từ biểu đồ so sánh (Hình 5.7) cho thấy khi áp dụng giải thuật CRMF vào 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).

Qua những thí nghiệm cho thấy ba giải thuật được cải tiến như: SocialMF, CRMF mang lại hiệu quả cao hơn các phương pháp cơ sở khác như: Students-kNNs, Courses-kNNs, Matrix factorization

**Hình 5.7:** Bảng so sánh độ lỗi RMSE của các giải thuật dự đoán

Các phương pháp để so sánh dùng trong thực nghiệm

- **Student-kNNs:** Phương pháp lọc cộng tác theo bộ nhớ dựa vào sinh viên tương tự

(luận án đã trình bày trong phần 3.2)

- **Course-kNNs:** Phương pháp lọc cộng tác theo bộ nhớ dựa vào môn học tương tự (luận án đã trình bày trong phần 3.3)
- **MF:** Phương pháp lọc cộng tác dựa vào mô hình phân rã ma trận (luận án đã trình bày trong phần 3.4)
- **Social-MF:** Phương pháp đề xuất sử dụng (giải thuật 5.1) Predicting Student Performance - SocialMF.
- **CRMF:** Phương pháp đề xuất sử dụng (giải thuật 5.2) Predicting Student Performance - CRRMF.

5.4 Tổng kết chương

Trong nghiên cứu này chúng tôi đã giới thiệu một hướng tiếp cận tích hợp mối quan hệ của người học vào dự đoán kết quả học tập của sinh viên, được minh họa thông qua kỹ thuật phân rã ma trận mạng xã hội (Social Matrix Factorization – SocialMF). Với tiếp cận này, ngoài những dữ liệu có sẵn, chúng ta có thể tận dụng thêm mối quan hệ giữa các người dùng này ở một hệ thống khác hoặc 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

Với sự phổ biến ngày càng tăng của mạng xã hội, đặc biệt là trong thế hệ học sinh và sinh viên trẻ, việc tích hợp dữ liệu mạng xã hội vào các hệ thống trợ giảng thông minh là một giải pháp thiết thực và khả thi. Tận dụng lượng thông tin khổng lồ được tạo ra thông qua các mạng xã hội, chẳng hạn như hoạt động, hành vi và sở thích của người dùng, có thể mang lại trải nghiệm học tập phù hợp và cá nhân hóa hơn cho sinh viên. Bằng cách tích hợp dữ liệu mạng xã hội vào các hệ thống trợ giảng thông minh, các nhà giáo dục có thể hiểu rõ hơn về phong cách và sở thích học tập của người học, đồng thời phát triển trải nghiệm học tập phù hợp để đáp ứng tốt hơn nhu cầu cá nhân của từng người học. Điều này có khả năng cải thiện đáng kể sự tham gia, động lực và kết quả học tập của sinh viên, đồng thời nâng cao chất lượng giáo dục tổng thể.

Bên cạnh đó chúng tôi còn đề xuất một hướng tiếp cận tích hợp mối liên quan của các môn học vào hệ trợ giảng thông minh, được minh họa thông qua kỹ thuật phân rã ma trận tích hợp mối liên quan các môn học (Courses Relationship Matrix Factorization –

CRMF). Với tiếp cận này, chúng tôi có thể tận dụng được thông tin chưa được khai thác như mô tả môn học, chuẩn đầu ra, kiến thức và kỹ năng cần đạt được của môn học để tích hợp vào xây dựng mô hình dự đoán, cải thiện được kết quả dự đoán.

Luận án đã có những đóng góp quan trọng trong việc nâng cao hiệu quả của dự đoán kết quả học tập, được thảo luận trong chương 3,4 và chương 5. Tuy nhiên, để có những đóng góp thiết thực, điều quan trọng là phải áp dụng kết quả của những nghiên cứu này trong các tình huống thực tế. Vì vậy, trong phần phụ lục của luận án có đề xuất xây dựng hệ thống gợi ý lựa chọn môn học tự chọn để lập kế hoạch học tập cho sinh viên nhằm nâng cao thành tích học tập của họ.

CHƯƠNG 6 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

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

Trên mặt lý thuyết, việc nghiên cứu các giải pháp gợi ý trong cố vấn học tập sẽ giúp cho các nhà nghiên cứu có thể hiểu sâu hơn về các phương pháp và kỹ thuật trong lĩnh vực này. Đồng thời, việc phát triển các giải pháp gợi ý trong cố vấn học tập cũng sẽ đóng góp vào việc giải quyết các vấn đề khó khăn liên quan đến việc xác định, đánh giá và cải thiện hiệu quả của các hoạt động học tập.

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, chúng tôi 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

Trên mặt thực tiễn, việc nghiên cứu các giải pháp gợi ý trong cố vấn học tập và xây dựng ứng dụng thử nghiệm sẽ giúp cho các giảng viên và sinh viên có thể sử dụng các công nghệ máy tính để cải thiện quá trình học tập. Các ứng dụng thử nghiệm được xây dựng dựa trên các giải pháp gợi ý sẽ giúp cho các sinh viên có thể lập kế hoạch học tập phù hợp với nhu cầu của mình, tăng cường hiệu quả học tập và giúp giảng viên theo dõi quá trình học tập của sinh viên một cách chính xác hơn.

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 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

Luận án còn có sự quan tâm đặc biệt từ cộng đồng nghiên cứu và nhà giáo dục, đó là một lý do quan trọng để nghiên cứu và phát triển các giải pháp gợi ý trong cố vấn học tập. Việc tạo ra các giải pháp gợi ý chất lượng và ứng dụng thực tế trong giáo dục sẽ

đóng góp vào việc nâng cao chất lượng giáo dục, giúp các sinh viên đạt được thành tích học tập cao hơn và phát triển sự nghiệp sau này.

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 đề tài. 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 của đề tài.

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.

CÁC CÔNG TRÌNH KHOA HỌC ĐÃ CÔNG BỐ

- [CT1] Methods for building course recommendation systems. In Proceedings of the 2016 International Conference on Knowledge and Systems Engineering (KSE 2016), pp.163-168, ISBN 978-1-4673-8929-7, **IEEE Xplore. Scopus-Indexed**
- [CT2] Toward integrating social networks into Intelligent Tutoring Systems. In Proceedings of the 2017 International Conference on Knowledge and Systems Engineering, (KSE 2017), pp.112-117, ISBN 978-1-5386-3576-6, **IEEE Xplore. Scopus-Indexed**
- [CT3] Integrating courses' relationship into predicting student performance, International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), 2020. ISSN: 2278–3091. **Scopus-Indexed.**
- [CT4] 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. **Springer, Cham. Scopus-Indexed**
- [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.

Tài liệu tham khảo

- [1] Viện NCGD, Trường ĐHSP TP.HCM, “Vai trò của cố vấn học tập trong đào tạo theo học chế tín chỉ tại các trường Cao đẳng - Đại học Việt Nam,” in *Vai trò của cố vấn học tập trong đào tạo theo học chế tín chỉ tại các trường Cao đẳng - Đại học Việt Nam*, 2014.
- [2] Nguyễn Thị Uyên và Nguyễn Minh Hiệp, “Dự đoán kết quả học tập của sinh viên bằng kỹ thuật khai phá dữ liệu,” *Tạp chí khoa học, Trường Đại học Vinh*, vol. 48, no. 3A/2019, pp. 68–73, 2019.
- [3] Lê Đức Thắng, Nguyễn Thái Nghe, Huỳnh Xuân Hiệp, và Trương Thị Hải, “Giải pháp hỗ trợ sinh viên lập kế hoạch học tập dựa trên tiếp cận tập thô,” in *Kỷ yếu Hội nghị Khoa học Quốc gia lần thứ IX “Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR’9)”*, 2016, pp. 151–158.
- [4] Lưu Hoài Sang, Trần Thanh Điện, Nguyễn Thanh Hải và Nguyễn Thái Nghe, “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*, vol. 56, no. 3A, pp. 20–28, 2020.
- [5] Đỗ Thị Liên, “Phát triển một số phương pháp xây dựng hệ thống tư vấn,” Ph.D. dissertation, Luận án tiến sĩ, Học Viện Bưu Chính Viễn Thông, 2020.
- [6] 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. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-73450-7_89
- [7] M. Cora Urdaneta-Ponte, A. Mendez-Zorrilla, I. Oleagordia-Ruiz, G. Kostopoulos, and S. Kotsiantis, “Recommendation Systems for Education: Systematic Review,” *Electronics 2021, Vol. 10, Page 1611*, vol. 10, no. 14, p. 1611, jul 2021. [Online]. Available: <https://www.mdpi.com/2079-9292/10/14/1611/htmlhttps://www.mdpi.com/2079-9292/10/14/1611>

- [8] 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*. Institute of Electrical and Electronics Engineers Inc., jan 2015, pp. 61–66.
- [9] C. Romero and S. Ventura, “Educational data mining and learning analytics: An updated survey,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 3, p. e1355, may 2020. [Online]. Available: <https://onlinelibrary.wiley.com/doi/full/10.1002/widm.1355><https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1355><https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.1355>
- [10] D. Buenaño-Fernández, D. Gil, and S. Luján-Mora, “Application of machine learning in predicting performance for computer engineering students: A case study,” *Sustainability (Switzerland)*, vol. 11, no. 10, may 2019.
- [11] R. Ghorbani and R. Ghousi, “Comparing Different Resampling Methods in Predicting Students’ Performance Using Machine Learning Techniques,” *IEEE Access*, vol. 8, pp. 67 899–67 911, 2020.
- [12] X. Li, X. Zhu, X. Zhu, Y. Ji, and X. Tang, “Student Academic Performance Prediction Using Deep Multi-source Behavior Sequential Network,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12084 LNAI, pp. 567–579, may 2020. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-47426-3_44
- [13] Y. Zhang, A. Ghandour, and V. Shestak, “Using Learning Analytics to Predict Students Performance in Moodle LMS,” *International Journal of Emerging Technologies in Learning (iJET)*, vol. 15, no. 20, pp. 102–115, oct 2020. [Online]. Available: <https://online-journals.org/index.php/i-jet/article/view/15915>
- [14] J. López-Zambrano, J. A. Lara, and C. Romero, “Towards Portability of Models for Predicting Students’ Final Performance in University Courses Starting from Moodle Logs,” *Applied Sciences 2020, Vol. 10, Page 354*, vol. 10, no. 1, p. 354, jan

2020. [Online]. Available: <https://www.mdpi.com/2076-3417/10/1/354/htmlhttps://www.mdpi.com/2076-3417/10/1/354>
- [15] R. Hasan, S. Palaniappan, A. R. A. Raziff, S. Mahmood, and K. U. Sarker, “Student Academic Performance Prediction by using Decision Tree Algorithm,” *2018 4th International Conference on Computer and Information Sciences: Revolutionising Digital Landscape for Sustainable Smart Society, ICCOINS 2018 - Proceedings*, oct 2018.
- [16] J. M. Luna, C. Castro, and C. Romero, “MDM tool: A data mining framework integrated into Moodle,” *Computer Applications in Engineering Education*, vol. 25, no. 1, pp. 90–102, jan 2017. [Online]. Available: <https://onlinelibrary.wiley.com/doi/full/10.1002/cae.21782https://onlinelibrary.wiley.com/doi/abs/10.1002/cae.21782https://onlinelibrary.wiley.com/doi/10.1002/cae.21782>
- [17] M. Liz-Domínguez, M. Caeiro-Rodríguez, M. Llamas-Nistal, and F. A. Mikic-Fonte, “Systematic Literature Review of Predictive Analysis Tools in Higher Education,” *Applied Sciences 2019, Vol. 9, Page 5569*, vol. 9, no. 24, p. 5569, dec 2019. [Online]. Available: <https://www.mdpi.com/2076-3417/9/24/5569/htmlhttps://www.mdpi.com/2076-3417/9/24/5569>
- [18] M. Llamas-Nistal, F. A. Mikic-Fonte, M. Caeiro-Rodriguez, and M. Liz-Dominguez, “Supporting intensive continuous assessment with BeA in a flipped classroom experience,” *IEEE Access*, vol. 7, pp. 150 022–150 036, 2019.
- [19] C. Yang, F. K. Chiang, Q. Cheng, and J. Ji, “Machine Learning-Based Student Modeling Methodology for Intelligent Tutoring Systems:,” <https://doi.org/10.1177/0735633120986256>, vol. 59, no. 6, pp. 1015–1035, jan 2021. [Online]. Available: <https://journals.sagepub.com/doi/abs/10.1177/0735633120986256>
- [20] N. Chaudhary and D. R. Chowdhury, “Data preprocessing for evaluation of recommendation models in E-commerce,” *Data*, vol. 4, no. 1, 2019.

- [21] I. Portugal, P. Alencar, and D. Cowan, “The use of machine learning algorithms in recommender systems: A systematic review,” *Expert Systems with Applications*, vol. 97, pp. 205–227, may 2018.
- [22] J. Wei, J. He, K. Chen, Y. Zhou, and Z. Tang, “Collaborative filtering and deep learning based recommendation system for cold start items,” *Expert Systems with Applications*, vol. 69, pp. 29–39, mar 2017.
- [23] M. C. Mihaescu and P. S. Popescu, “Review on publicly available datasets for educational data mining,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 3, p. e1403, may 2021. [Online]. Available: <https://onlinelibrary.wiley.com/doi/full/10.1002/widm.1403><https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1403><https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.1403>
- [24] F. O. Isinkaye, Y. O. Folajimi, and B. A. Ojokoh, “Recommendation systems: Principles, methods and evaluation,” pp. 261–273, nov 2015.
- [25] R. Chen, Q. Hua, B. Wang, M. Zheng, W. Guan, X. Ji, Q. Gao, and X. Kong, “A Novel Social Recommendation Method Fusing User’s Social Status and Homophily Based on Matrix Factorization Techniques,” *IEEE Access*, vol. 7, pp. 18 783–18 798, 2019.
- [26] M. Fu, H. Qu, D. Moges, and L. Lu, “Attention based collaborative filtering,” *Neurocomputing*, vol. 311, pp. 88–98, oct 2018.
- [27] P. Lops, D. Jannach, C. Musto, T. Bogers, and M. Koolen, “Trends in content-based recommendation: Preface to the special issue on Recommender systems based on rich item descriptions,” *User Modeling and User-Adapted Interaction*, vol. 29, no. 2, pp. 239–249, apr 2019. [Online]. Available: <https://link.springer.com/article/10.1007/s11257-019-09231-w>
- [28] R. Du, J. Lu, and H. Cai, “Double Regularization Matrix Factorization Recommendation Algorithm,” *IEEE Access*, vol. 7, pp. 139 668–139 677, 2019.
- [29] M. Hahsler, “recommenderlab: An R Framework for Developing and Testing Recommendation Algorithms,” *arXiv e-prints*, p. arXiv:2205.12371, May 2022.

- [30] J. Lu, D. Wu, M. Mao, W. Wang, and G. Zhang, “Recommender system application developments: A survey,” *Decision Support Systems*, vol. 74, pp. 12–32, jun 2015.
- [31] S. Zhang, L. Yao, A. Sun, and Y. Tay, “Deep learning based recommender system: A survey and new perspectives,” feb 2019.
- [32] 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,” *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 8, pp. 711–721, 2020. [Online]. Available: www.ijacsa.thesai.org
- [33] M. Zhang, B. Hu, C. Shi, B. Wu, and B. Wang, “Matrix Factorization Meets Social Network Embedding for Rating Prediction,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10987 LNCS, pp. 121–129, jul 2018. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-96890-2_10
- [34] J. S. Gil, A. J. P. Delima, and R. N. Vilchez, “Predicting students’ dropout indicators in public school using data mining approaches,” *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 1, pp. 774–778, 2020.
- [35] J. Kabathova and M. Drlik, “Towards Predicting Student’s Dropout in University Courses Using Different Machine Learning Techniques,” *Applied Sciences 2021, Vol. 11, Page 3130*, vol. 11, no. 7, p. 3130, apr 2021. [Online]. Available: <https://www.mdpi.com/2076-3417/11/7/3130/htmlhttps://www.mdpi.com/2076-3417/11/7/3130>
- [36] 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,” *International Journal of Innovative Computing, Information and Control*, vol. 14, no. 6, pp. 2129–2140, dec 2018. [Online]. Available: <https://researchoutput.ncku.edu.tw/en/publications/the-predictive-modeling-for-learning-student-results-based-on-seq>
- [37] 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,” *Emerging Science Journal*, vol. 2, no. 2, pp. 100–107, may 2018. [Online]. Available: <https://www.ijournalse.org/index.php/ESJ/article/view/73>
- [38] 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 Systems*, vol. 194, p. 105385, apr 2020.
- [39] B. C. Madeira, T. Tasci, and N. Celebi, “Prediction of Student Performance Using Rough Set Theory And Backpropagation Neural Networks,” *European Scientific Journal, ESJ*, vol. 17, no. 7, pp. 1–1, feb 2021. [Online]. Available: <https://eujournal.org/index.php/esj/article/view/14028>
- [40] 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,” *Proceedings - 2nd East Indonesia Conference on Computer and Information Technology: Internet of Things for Industry, EIconCIT 2018*, pp. 78–81, nov 2018.
- [41] X. Du, J. Yang, J. L. Hung, and B. Shelton, “Educational data mining: a systematic review of research and emerging trends,” *Information Discovery and Delivery*, vol. 48, no. 4, pp. 225–236, oct 2020.
- [42] P. Sharma, . M. Harkishan, and M. Harkishan, “Designing an intelligent tutoring system for computer programing in the Pacific,” *Education and Information Technologies 2022*, pp. 1–13, jan 2022. [Online]. Available: <https://link.springer.com/article/10.1007/s10639-021-10882-9>
- [43] R. Lara-Cabrera, A. González-Prieto, and F. Ortega, “Deep Matrix Factorization Approach for Collaborative Filtering Recommender Systems,” *Applied Sciences 2020, Vol. 10, Page 4926*, vol. 10, no. 14, p. 4926, jul 2020. [Online]. Available: <https://www.mdpi.com/2076-3417/10/14/4926/htmhttps://www.mdpi.com/2076-3417/10/14/4926>
- [44] T. T. Dien, N. Thanh-Hai, and N. T. Nghe, “Deep Matrix Factorization for Learning Resources Recommendation,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in*

Bioinformatics), vol. 12876 LNAI, pp. 167–179, sep 2021. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-88081-1_13

- [45] C. Xu, “A novel recommendation method based on social network using matrix factorization technique,” *Information Processing Management*, vol. 54, no. 3, pp. 463–474, may 2018.
- [46] Y. Yu, C. Wang, H. Wang, and Y. Gao, “Attributes coupling based matrix factorization for item recommendation,” *Applied Intelligence*, vol. 46, no. 3, pp. 521–533, apr 2017. [Online]. Available: <https://dl.acm.org/doi/abs/10.1007/s10489-016-0841-8>
- [47] F. Ricci, L. Rokach, and B. Shapira, “Introduction to Recommender Systems Handbook,” in *Recommender Systems Handbook*. Springer US, 2011, pp. 1–35.
- [48] J. Xu, J.-W. Karaleise, and S. Li, “The development, status and trends of recommender systems: a comprehensive and critical literature review,” *undefined*, 2014.
- [49] M. Arevalillo-Herráez, D. Arnau, and L. Marco-Giménez, “Domain-specific knowledge representation and inference engine for an intelligent tutoring system,” *Knowledge-Based Systems*, vol. 49, pp. 97–105, sep 2013.
- [50] D. Dermeval, R. Paiva, I. I. Bittencourt, J. Vassileva, and D. Borges, “Authoring Tools for Designing Intelligent Tutoring Systems: a Systematic Review of the Literature,” *International Journal of Artificial Intelligence in Education*, vol. 28, no. 3, pp. 336–384, sep 2018. [Online]. Available: <https://link.springer.com/article/10.1007/s40593-017-0157-9>
- [51] M. Yani and A. A. Krisnadhi, “Challenges, Techniques, and Trends of Simple Knowledge Graph Question Answering: A Survey,” *Information 2021, Vol. 12, Page 271*, vol. 12, no. 7, p. 271, jun 2021. [Online]. Available: <https://www.mdpi.com/2078-2489/12/7/271/htmhttps://www.mdpi.com/2078-2489/12/7/271>
- [52] B. D. Nye, A. C. Graesser, and X. Hu, “AutoTutor and Family: A Review of 17 Years of Natural Language Tutoring.” *Grantee Submission*, vol. 24, no. 4, pp. 427–469, oct 2014.

- [53] Y. Y. Wang, A. F. Lai, R. K. Shen, C. Y. Yang, V. R. Shen, and Y. H. Chu, “Modeling and verification of an intelligent tutoring system based on Petri net theory,” *Mathematical biosciences and engineering : MBE*, vol. 16, no. 5, pp. 4947–4975, 2019. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/31499699/>
- [54] Q. Wang, S. Jing, D. Joyner, L. Wilcox, H. Li, T. Plötz, and B. Disalvo, “Sensing Affect to Empower Students: Learner Perspectives on Affect-Sensitive Technology in Large Educational Contexts,” *L@S 2020 - Proceedings of the 7th ACM Conference on Learning @ Scale*, pp. 63–76, aug 2020.
- [55] N. Khodeir, “Student Modeling Using Educational Data Mining Techniques,” *ACC-S/PEIT 2019 - 2019 6th International Conference on Advanced Control Circuits and Systems and 2019 5th International Conference on New Paradigms in Electronics and Information Technology*, pp. 7–14, nov 2019.
- [56] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, “Recommender systems survey,” *Knowledge-Based Systems*, vol. 46, pp. 109–132, jul 2013.
- [57] X. Su and T. M. Khoshgoftaar, “A Survey of Collaborative Filtering Techniques,” *Advances in Artificial Intelligence*, vol. 2009, pp. 1–19, oct 2009.
- [58] K. Haruna, M. A. Ismail, S. Suhendroyono, D. Damiasih, A. C. Pierewan, H. Chiroma, and T. Herawan, “Context-aware recommender system: A review of recent developmental process and future research direction,” dec 2017.
- [59] R. Burke, “Hybrid recommender systems: Survey and experiments,” *User Modelling and User-Adapted Interaction*, vol. 12, no. 4, pp. 331–370, 2002.
- [60] T. Mohammadpour, A. M. Bidgoli, R. Enayatifar, and H. H. S. Javadi, “Efficient clustering in collaborative filtering recommender system: Hybrid method based on genetic algorithm and gravitational emulation local search algorithm,” *Genomics*, vol. 111, no. 6, pp. 1902–1912, dec 2019.
- [61] L. Ren and W. Wang, “An SVM-based collaborative filtering approach for Top-N web services recommendation,” *Future Generation Computer Systems*, vol. 78, pp. 531–543, jan 2018.

- [62] S. Linda and K. K. Bharadwaj, "A Decision Tree Based Context-Aware Recommender System," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11278 LNCS, pp. 293–305, dec 2018. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-04021-5_27
- [63] M. Fu, H. Qu, Z. Yi, L. Lu, and Y. Liu, "A Novel Deep Learning-Based Collaborative Filtering Model for Recommendation System," *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 1084–1096, mar 2019.
- [64] X. Dai, F. Li, X. Li, and H. Liang, "A Novel Recommender System using Hidden Bayesian Probabilistic Model based Collaborative Filtering," *Proceedings of the International Joint Conference on Neural Networks*, vol. 2019-July, jul 2019.
- [65] M. Nilashi, O. Ibrahim, and K. Bagherifard, "A recommender system based on collaborative filtering using ontology and dimensionality reduction techniques," *Expert Systems with Applications*, vol. 92, pp. 507–520, feb 2018.
- [66] H. Mohana and D. M. Suriakala, "A Study on Ontology Based Collaborative Filtering Recommendation Algorithms in E-Commerce Applications," *IOSR Journal of Computer Engineering*, vol. 19, no. 04, pp. 14–19, jun 2017.
- [67] E. Çano and M. Morisio, "Hybrid recommender systems: A systematic literature review," *Intelligent Data Analysis*, vol. 21, no. 6, pp. 1487–1524, jan 2017.
- [68] R. Nzeh, R. Nzeh, N. J. Ezeora, U. Izuchukwu, and U. B. Chimezie, "Optimization Of An Online Store Price Recommendation System Using Hybrid," *International Journal of Progressive Sciences and Technologies*, vol. 27, no. 2, pp. 431–440, jul 2021. [Online]. Available: <https://ijpsat.ijsh-t-journals.org/index.php/ijpsat/article/view/3313>
- [69] G. Geetha, M. Safa, C. Fancy, and D. Saranya, "A Hybrid Approach using Collaborative filtering and Content based Filtering for Recommender System," *Journal of Physics: Conference Series*, vol. 1000, no. 1, p. 012101, apr 2018. [Online]. Available: <https://iopscience.iop.org/article/10.1088/1742-6596/1000/1/012101><https://iopscience.iop.org/article/10.1088/1742-6596/1000/1/012101/meta>

- [70] T. Patikorn, R. S. Baker, and N. T. Heffernan, “ASSISTments Longitudinal Data Mining Competition Special Issue: A Preface,” *Journal of Educational Data Mining*, vol. 12, no. 2, pp. i–xi, aug 2020. [Online]. Available: <https://jedm.educationaldatamining.org/index.php/JEDM/article/view/486https://jedm.educationaldatamining.org>
- [71] N. T. Nghe, P. Janecek, and P. Haddawy, “A comparative analysis of techniques for predicting academic performance,” in *2007 37th Annual Frontiers In Education Conference - Global Engineering: Knowledge Without Borders, Opportunities Without Passports*, 2007, pp. T2G–7–T2G–12.
- [72] G. Adomavicius and A. Tuzhilin, “Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions,” pp. 734–749, jun 2005. [Online]. Available: <https://experts.umn.edu/en/publications/toward-the-next-generation-of-recommender-systems-a-survey-of-the>
- [73] Y. Koren, R. Bell, and C. Volinsky, “Matrix factorization techniques for recommender systems,” *Computer*, vol. 42, no. 8, pp. 30–37, 2009.
- [74] L. A. Dieker, C. E. Hughes, M. C. Hynes, and C. Straub, “Using Simulated Virtual Environments to Improve Teacher Performance,” *School-University Partnerships*, 2017.
- [75] F. Ortega, J. Mayor, D. López-Fernández, and R. Lara-Cabrera, “Cf4j 2.0: Adapting collaborative filtering for java to new challenges of collaborative filtering based recommender systems,” *Knowledge-Based Systems*, vol. 215, p. 106629, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0950705120307589>
- [76] M. D. Ekstrand, J. T. Riedl, and J. A. Konstan, “Collaborative filtering recommender systems,” *Foundations and Trends® in Human–Computer Interaction*, vol. 4, no. 2, pp. 81–173, 2011. [Online]. Available: <http://dx.doi.org/10.1561/11000000009>
- [77] J. Liu, G. Xu, and Z. Ying, “Theory of self-learning Q-matrix,” <https://doi.org/10.3150/12-BEJ430>, vol. 19, no. 5A, pp. 1790–

1817, nov 2013. [Online]. Available: <https://projecteuclid.org/journals/bernoulli/volume-19/issue-5A/Theory-of-self-learning-Q-matrix/10.3150/12-BEJ430.full>
<https://projecteuclid.org/journals/bernoulli/volume-19/issue-5A/Theory-of-self-learning-Q-matrix/10.3150/12-BEJ430.short>

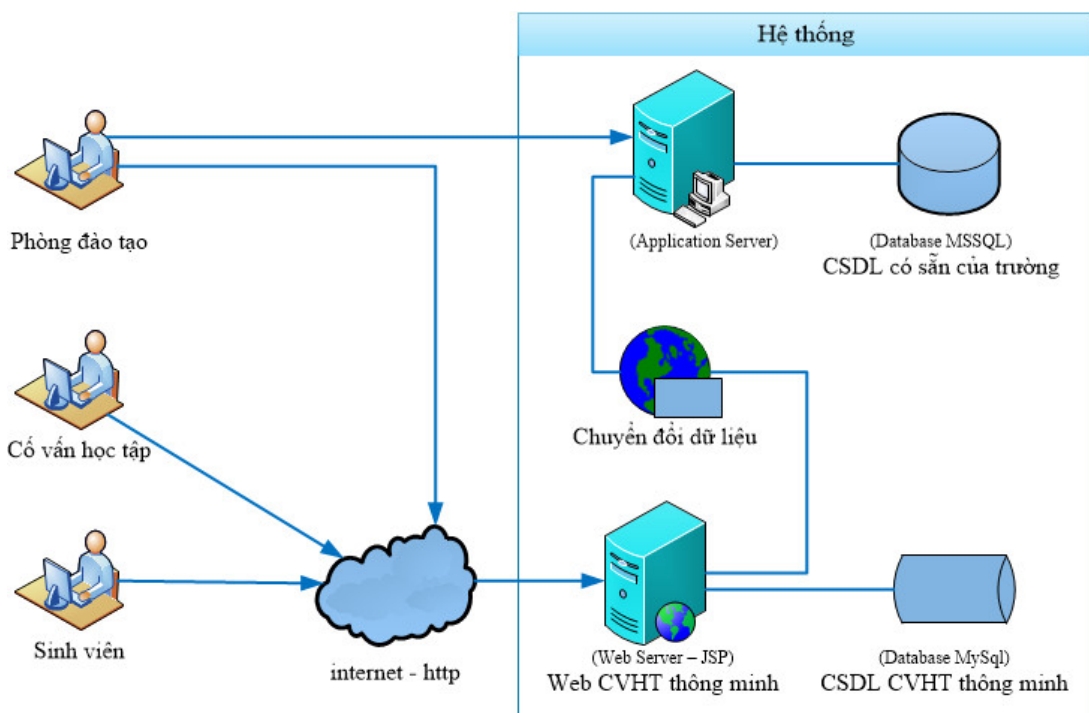
- [78] J. R. Anderson, A. T. Corbett, K. R. Koedinger, and R. Pelletier, “An Intelligent Tutoring System for Learning Android Applications UI Development,” *Journal of the Learning Sciences*, vol. 4, no. 2, pp. 167–207, jan 2018. [Online]. Available: <https://papers.ssrn.com/abstract=3104365>
- [79] J. Paladines and J. Ramírez, “A systematic literature review of intelligent tutoring systems with dialogue in natural language,” *IEEE Access*, vol. 8, pp. 164 246–164 267, 2020.

PHỤ LỤC

Tổng quan về ứng dụng thử nghiệm

Do dữ liệu lớn và thời gian xử lý dữ liệu, chạy giải thuật dự đoán đòi hỏi tương đối lâu, bên cạnh hệ thống chứa những thông tin nhạy cảm (điểm số), nên chúng tôi đề xuất phát triển trên nền ứng dụng desktop. Trong khi đó, hệ thống gợi ý đòi hỏi sinh viên phải được tư vấn nhanh chóng, tiện lợi, mọi lúc mọi nơi và phục vụ nhiều sinh viên trong khoảng thời gian nhất định nên chúng tôi đề xuất xây dựng trên hệ thống trên nền web.

Hình 1 thể hiện kiến trúc tổng quan hệ thống thử nghiệm. Hệ thống đề xuất gồm ba ứng dụng chính: ứng dụng desktop, ứng dụng web, và ứng dụng chuyển đổi dữ liệu. Về cơ sở dữ liệu chúng sẽ được lưu trữ ở hai nơi để đảm bảo tính toàn vẹn của cơ sở dữ liệu sẵn có của trường đang hoạt động.



Hình 1: Hình kiến trúc hệ thống triển khai

Phân tích thiết kế hệ thống

Để xây dựng hệ thống dự đoán và gợi ý, đầu tiên chúng tôi cần thiết kế một cơ sở dữ liệu để lưu trữ dữ liệu kế hoạch học tập và điểm số của sinh viên phù hợp với đào tạo

theo tín chỉ. Do mỗi trường đại học có mỗi cách quản lý khác nhau. Nên chúng tôi xin trình bày hệ thống được đề xuất chỉ đáp ứng một số vấn đề cơ bản trong hệ thống đào tạo tín chỉ.

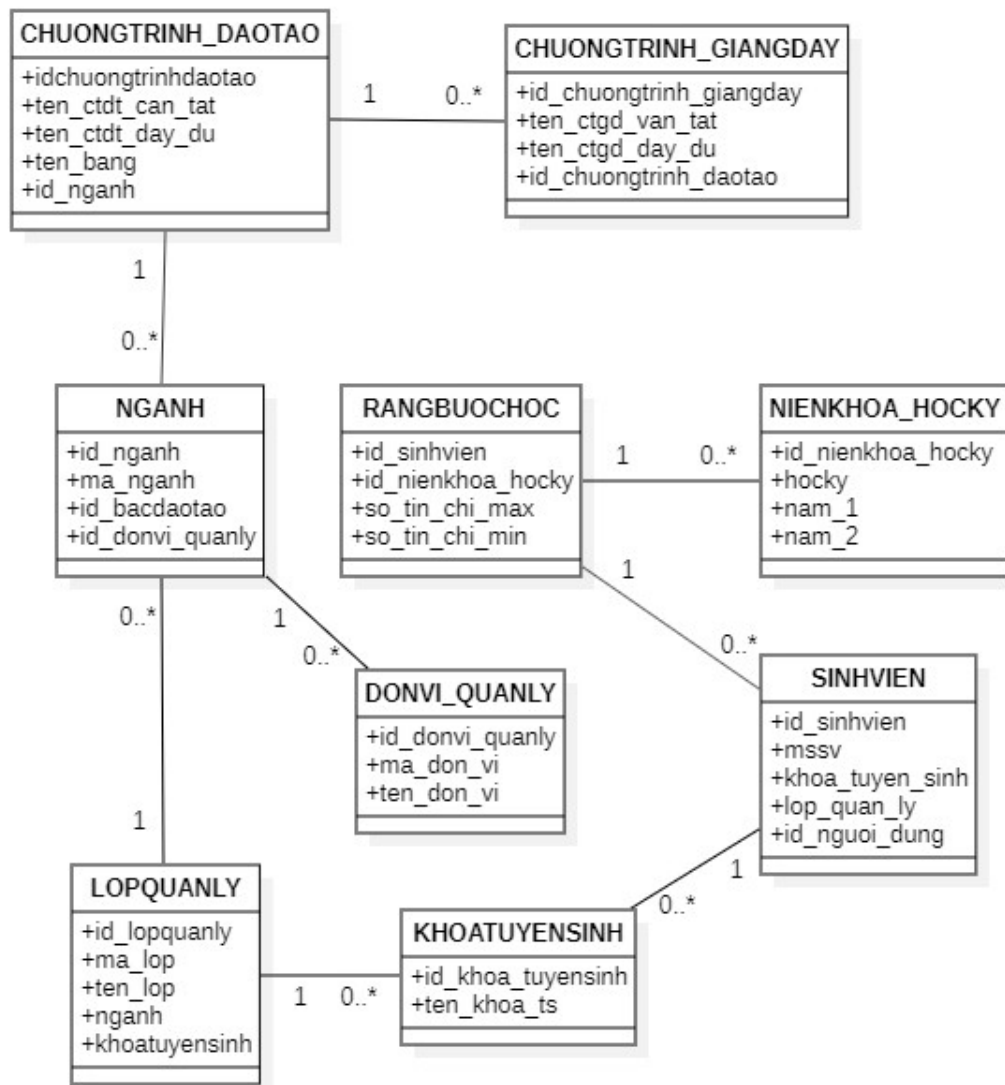
Cơ sở dữ liệu chương trình đào tạo được thiết kế đơn giản chỉ liên quan đến chức năng gợi ý môn học nên chúng không chứa những thông tin quản lý khác. Hệ thống tập trung xử lý ba nhóm dữ liệu như sau: sinh viên, môn học và điểm số (Users, Items, Ratings).

- Nhóm môn học (Items) bao gồm những bảng dữ liệu như sau: kế hoạch học tập mẫu theo từng học kỳ, môn học tiên quyết, môn học bắt buộc, nhóm môn tự chọn, môn học tự chọn, ...) cùng với những ràng buộc toàn vẹn.
- Nhóm sinh viên (Users) gồm có: sinh viên, lớp, ngành học, khóa học.
- Nhóm điểm số (Ratings) gồm có: sinh viên, môn học, điểm hệ số 10, điểm hệ số 4 tương đương với đánh giá năng lực của sinh viên, học kỳ, năm học.

Bên cạnh việc xây dựng hệ thống gợi ý lựa chọn môn học, cơ sở dữ liệu này phải được thiết kế phù hợp theo chương trình đào tạo mà nhà trường đang quản lý.

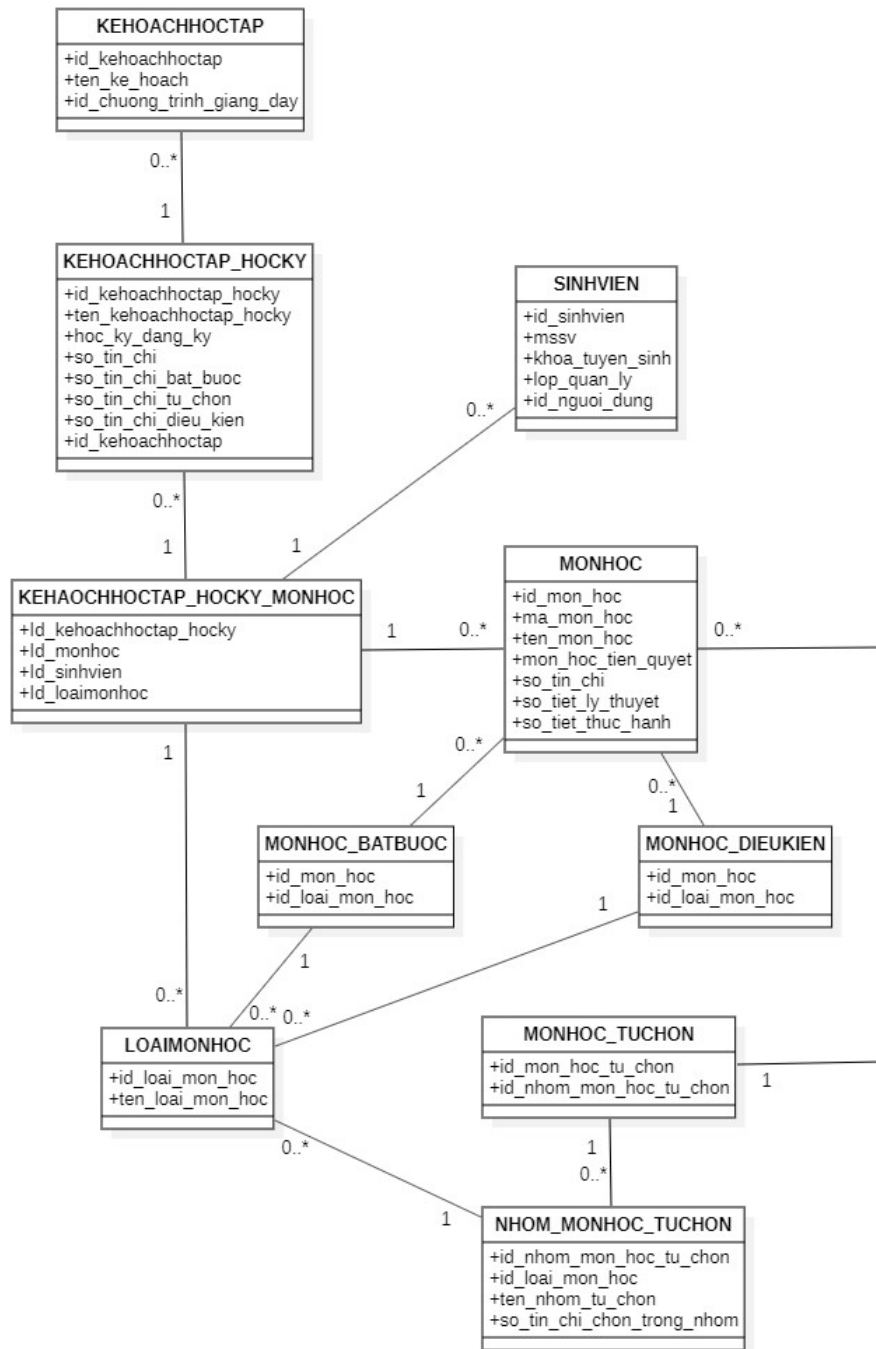
Theo định dạng của giải thuật MF thì chúng tôi cần ba thành phần dữ liệu: dữ liệu người dùng (user), dữ liệu đối tượng môn học (item), dữ liệu đánh giá hay ở đây là điểm số (ratings). Bên cạnh đó, giải thuật tích hợp mạng xã hội (Social Network MF) thì dữ liệu dùng để dự đoán cần thêm thông tin lớp học của sinh viên; giải thuật tích hợp mối liên quan giữa các môn học (Courses Relationship MF) thì cần thêm thông tin kiến thức của môn học.

Sơ đồ lớp liên quan đến nhóm sinh viên (users) của hệ thống được trình bày trong (Hình (2)). Ở nhóm dữ liệu này, sinh viên có thể hiện mối quan hệ với nhau như lớp quản lý, ngành tuyển sinh, khóa học. Từ những dữ liệu này, ta có thể khai thác thông tin thêm để có thể tích hợp và nâng cao hiệu quả dự đoán kết quả học tập của sinh viên.



Hình 2: Sơ đồ lớp dữ liệu liên quan đến nhóm sinh viên (users)

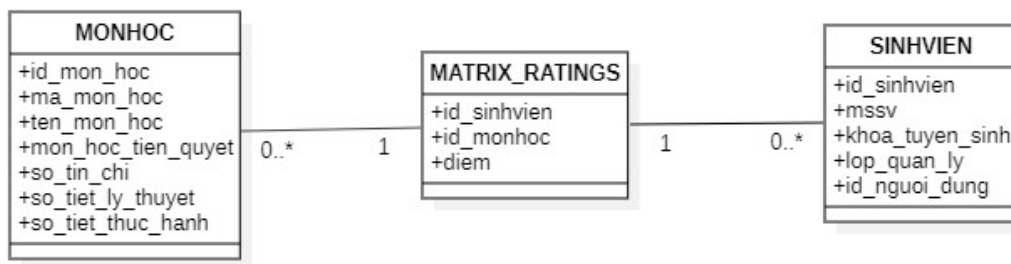
Sơ đồ lớp liên quan đến nhóm môn học (items) của hệ thống được trình bày trong (Hình (3)). Ở nhóm dữ liệu này, các môn học có ràng buộc về chương trình đào tạo như: các môn học bắt buộc, các môn học tiên quyết, môn học tự chọn và môn học song hành, bên cạnh đó chúng còn liên kết với học kỳ và kế hoạch học tập của sinh viên. Thiết kế dữ liệu này nhằm giúp hệ thống đưa ra gợi ý lựa chọn môn học bởi cách thực hiện đơn giản là: lọc ra các điểm dự đoán cao nhất của danh sách môn học thỏa mãn những ràng buộc này.



Hình 3: Sơ đồ lớp liên quan đến dữ liệu môn học (items)

Sơ đồ lớp liên quan đến điểm số (Ratings) của hệ thống được thể hiện (Hình (4)) được thiết kế để phục vụ sẵn sàng cho ứng dụng gợi ý, mà không phải thực hiện tính toán của các thuật toán khai phá dữ liệu đòi hỏi máy chủ phải đủ mạnh.

Sau khi thiết kế được cơ sở dữ liệu lưu trữ, để việc thực hiện các giải thuật dự đoán, các dữ liệu này cần được chuẩn bị và tiền xử lý. Quá trình tiền xử lý được thực hiện như sau: tiến hành xử lý dữ liệu sai định dạng hoặc rỗng, kế tiếp là chuyển dữ liệu về



Hình 4: Sơ đồ lớp liên quan đến dữ liệu điểm (ratings)

định dạng theo yêu cầu của các giải thuật gợi ý là (user, item, ratings). Quá trình chuyển đổi gồm ba bước cơ bản: từ bảng sinh viên chuyển thành danh sách *user*, bảng môn học chuyển thành danh sách *item*, và cuối cùng là bảng điểm sẽ chuyển thành ma trận đánh giá *ratings*.

Bên cạnh đó, dữ liệu dùng để tích hợp nhằm nâng cao hiệu quả của dự đoán là: chuyển thuộc tính lớp học của bảng sinh viên thành ma trận mối quan hệ người dùng (Social Matrix), kiến thức của môn học chuyển thành ma trận liên quan của các môn học (Courses Relationship Matrix), những ma trận này là ma trận nhị phân đối xứng.

Các chức năng cơ bản trong hệ thống

Hệ thống website lập kế hoạch học tập bao gồm hai phần chính: quản lý chương trình đào tạo và quản lý kế hoạch học tập.

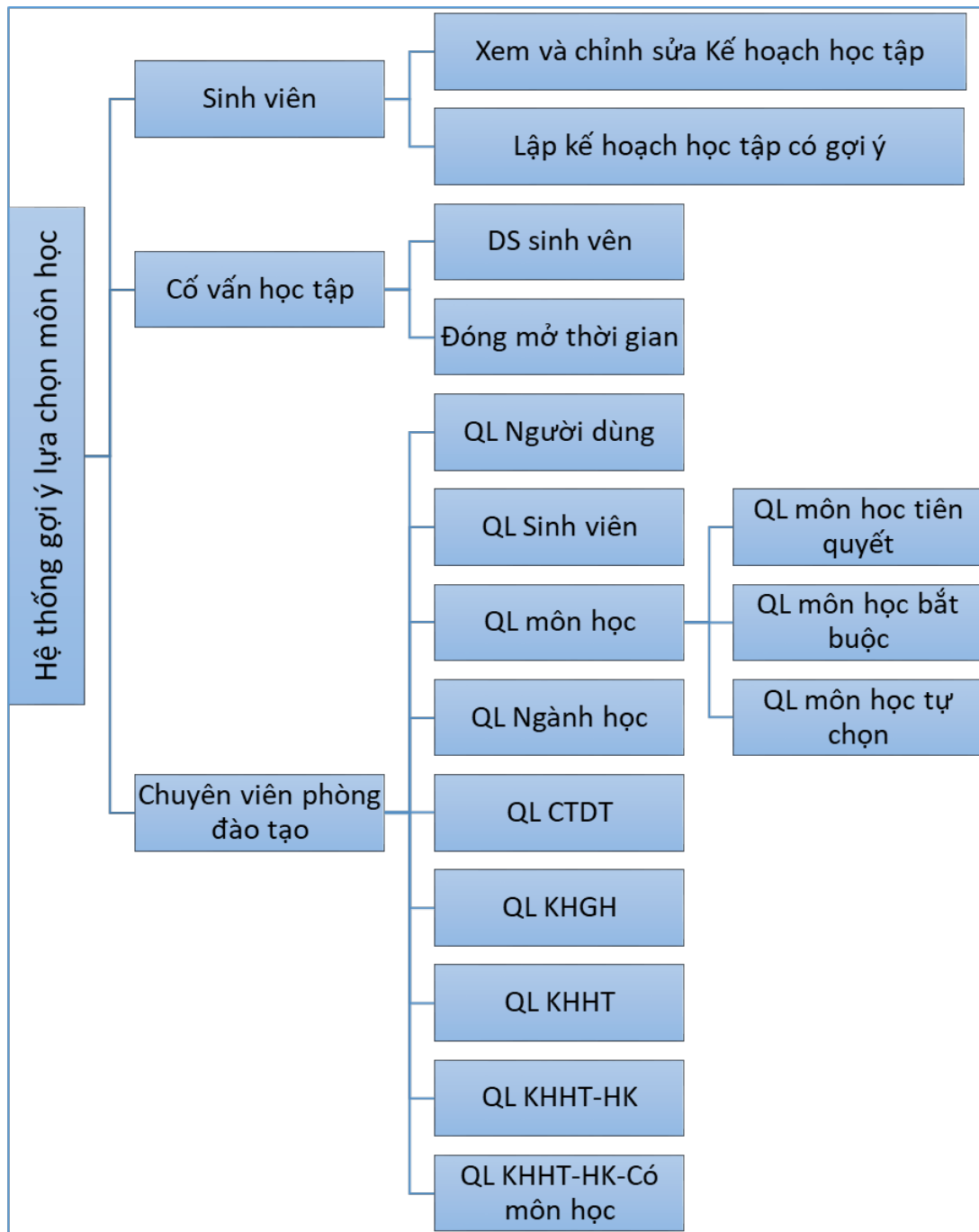
Quản lý chương trình đào tạo bao gồm: quản lý ngành, lớp tổ chức, khóa tuyển sinh, môn học, môn học bắt buộc, môn học tiên quyết, nhóm môn học tự chọn, môn học tự chọn, chương trình đào tạo, chương trình giảng dạy, kế hoạch học tập, kế hoạch học tập mẫu cho từng học kỳ có môn học.

Chức năng quản lý kế hoạch học tập: khi chuyên viên phòng đào tạo đã tạo kế hoạch học tập mẫu từng học kỳ có những môn học đã định sẵn mà phòng đào tạo và khoa đã thiết kế, thì hệ thống sẽ đặt kế hoạch học tập mẫu này cho tất cả sinh viên thuộc ngành và khóa tuyển sinh đó. Khi sinh viên có nhu cầu thay đổi môn học thì tự điều chỉnh trong kế hoạch học tập của họ.

Để xây dựng hệ thống gợi ý lựa chọn môn học phù hợp, chúng tôi đã tích hợp hệ thống gợi ý vào ngay chức năng lập kế hoạch học tập của người dùng sinh viên. Khi sinh viên đăng nhập vào hệ thống sẽ thấy được kế hoạch học tập mẫu theo khóa tuyển sinh và

ngành học của họ. Bên cạnh đó, việc gợi ý sẽ được tích hợp vào chức năng xem và điều chỉnh kế hoạch học tập này.

Sơ đồ chức năng của hệ thống ứng dụng nền web được trình bày trong Hình (5)



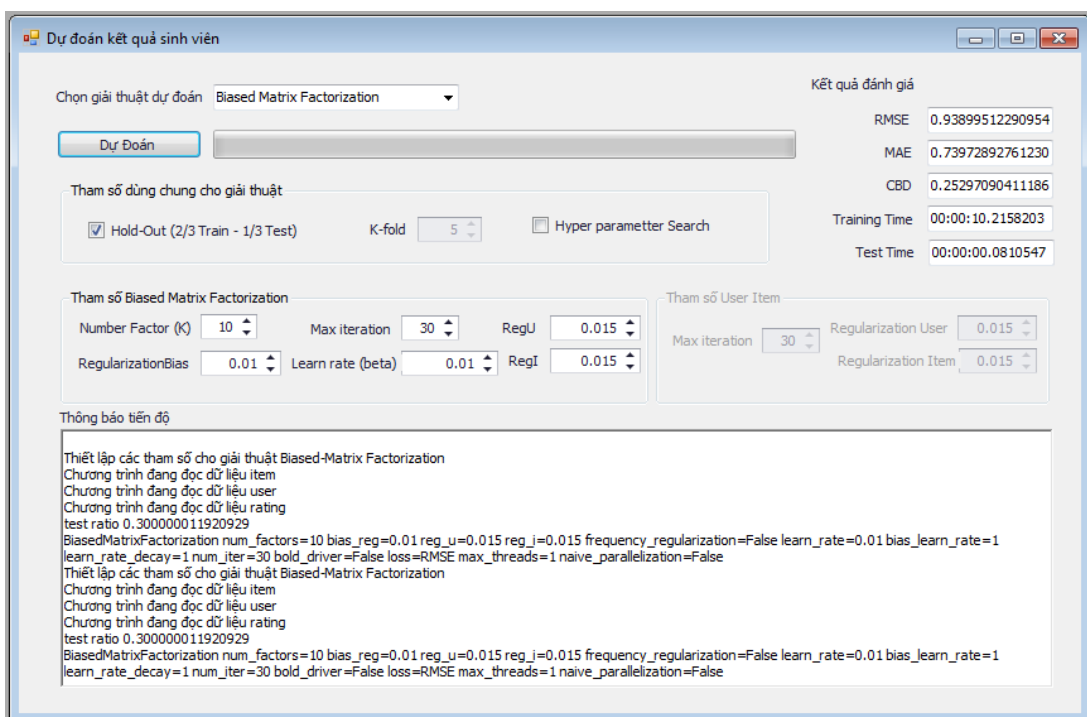
Hình 5: Sơ đồ chức năng của hệ thống nền web

Giao diện hệ thống

Như đã được trình bày phần đầu chương, kiến trúc hệ thống được thiết kế thành hai thành phần chính: (1) Hệ thống dự đoán (Ứng dụng desktop) thực hiện chức năng dự đoán tất cả điểm của sinh viên trên tất cả các môn học mà sinh viên đó chưa học; (2) Hệ

thống gợi ý (Ứng dụng web) để lập và xem kế hoạch học tập của sinh viên, ở đây sinh viên sẽ thấy được những 'gợi ý tự động' từ hệ thống với những thông tin dự đoán từ hệ thống dự đoán đã thực hiện và chuyển qua.

Tại ứng dụng desktop (hệ thống dự đoán), hệ thống được nhận dữ liệu đầu vào từ hệ thống web hoặc hệ thống quản lý điểm sinh viên sẵn có tại các trường đại học, thông qua một công cụ chuyển đổi dữ liệu. Sau khi dữ liệu được chuẩn hóa và tiền xử lý, hệ thống dự đoán sẽ tiến hành dự đoán tất cả các điểm của sinh viên theo một giải thuật toán tốt nhất với độ lỗi thấp (do quản trị viên hoặc quản lý lựa chọn). Ví dụ như (Hình 6) thể hiện ứng dụng đang trong quá trình dự đoán với giải thuật Biased-MF.



Hình 6: Giao diện ứng dụng desktop chức năng dự đoán kết quả học tập sinh viên

Tại ứng dụng web (hệ thống gợi ý), hệ thống này là dạng hệ thống quản lý kế hoạch tập của sinh viên trên nền web. Sinh viên có thể đăng nhập vào hệ thống để xem và lập kế hoạch học tập trước khi đăng ký học phần chính thức. Điểm hay ở hệ thống này đã nhận được các "điểm số đã dự đoán" từ hệ thống dự đoán, khi đó dựa trên ràng buộc như: số môn học tự chọn trong 1 học kỳ, môn học trước, và những môn cần học trong học kỳ này,...và điểm dự đoán sẽ tạo ra gợi ý đơn giản với số điểm dự đoán cao nhất.

Ví dụ như (Hình 7) thể hiện hệ thống web xem kế hoạch học tập có gợi ý tại một số môn học tự chọn. Tuy nhiên quyền chọn đăng ký môn học là quyền của sinh viên.

Xem KHHT toàn khóa
Cập nhật học phần phải học
Danh sách học phần
Cập nhật năm học - học kỳ

Trang chủ
Giới thiệu hệ thống
Giới thiệu giải thuật BMF
Hướng dẫn sử dụng - SV
Hướng dẫn sử dụng - Cán bộ
Tài liệu tham khảo
Các bài báo liên quan
Liên hệ



KẾ HOẠCH HỌC TẬP TOÀN KHÓA

Mã sinh viên: sv001 - Họ và tên: **Huỳnh Lý Thanh Nhân** - Lớp: **Kỹ Thuật Phần Mềm K32** - Khóa: **32**

STT	Mã Môn	Tên học phần	Loại HP	Số TC	Dự đoán & Gợi ý	Số tiết LT	Số tiết TH	HP tiên quyết	Năm học	Học kỳ
Năm học 2010-2011 Học kỳ 1										
1	QP001	Giáo dục quốc phòng	Điều kiện	6		115	50		2010-2011	1
2	TN001	Vi-Tích phần A1	Bắt Buộc	3		45	0		2010-2011	1
3	CT002	TT.Tin học căn bản	Bắt Buộc	2		0	60		2010-2011	1
4	CT001	Tin học căn bản	Bắt Buộc	1		15	0		2010-2011	1
Năm học 2010-2011 Học kỳ 2										
5	CT101	Lập trình căn bản A	Bắt Buộc	4		30	60		2010-2011	2
6	TN012	Đại số tuyến tính & Hình học	Bắt Buộc	4		60	0		2010-2011	2
7	TN010	Xác suất thống kê	Bắt Buộc	3		45	0		2010-2011	2
8	CT801	Anh văn căn bản 1	Tự Chọn	4	2.9	60	0		2010-2011	2
9	XH004	Pháp văn căn bản 1	Chọn 3 tín chỉ	3	3.6	45	0		2010-2011	2
Năm học 2010-2011 Học kỳ Hè										
10	CT103	Cấu trúc dữ liệu	Bắt Buộc	4		45	30		2010-2011	Hè
11	CT119	Toán rời rạc 2	Bắt Buộc	3		45	0	CT101	2010-2011	Hè
12	TN002	Vi-Tích phần A2	Bắt Buộc	4		60	0	TN001	2010-2011	Hè
13	XH005	Pháp văn căn bản 2	Tự Chọn	3	3.9	45	0	XH004	2010-2011	Hè
14	CT802	Anh văn căn bản 2	Chọn 3 tín chỉ	3	4.0	45	0	CT801	2010-2011	Hè
Năm học 2011-2012 Học kỳ 1										
15	TC100	Giáo dục thể chất 1	Điều kiện	1		0	45		2011-2012	1
16	ML006	Tư tưởng Hồ Chí Minh	Bắt Buộc	2		30	0	ML010	2011-2012	1
17	CT120	Phân tích & thiết kế thuật toán	Bắt Buộc	2		30	0	CT119	2011-2012	1
18	CT104	Kiến trúc máy tính	Bắt Buộc	2		30	0		2011-2012	1
19	CT114	Lập trình hướng đối tượng C++	Bắt Buộc	3		30	30	CT101	2011-2012	1
20	CT803	Anh văn căn bản 3	Tự Chọn	3	2.4	45	0	CT802	2011-2012	1
21	XH006	Pháp văn căn bản 3	Chọn 3 tín chỉ	4	3.5	60	0	XH005	2011-2012	1
22	CT123	Quy hoạch tuyến tính - CNTT	Tự Chọn Chọn 2 tín chỉ	2	1.5	30	0		2011-2012	1
23	CT126	Lý thuyết xếp hàng		2	2.4	30	0		2011-2012	1
24	CT124	Phương pháp tính - CNTT		2	2.6	30	0		2011-2012	1
25	CT125	Mô phỏng		2	2.9	30	0		2011-2012	1
26	CT127	Lý thuyết thông tin		2	3.1	30	0		2011-2012	1

Hình 7: Giao diện xem kế hoạch học tập của ứng dụng web

Tổng kết

Phụ lục đã đề xuất kiến trúc, mô hình để xây dựng các ứng dụng thực hiện việc dự đoán kết quả học tập của sinh viên, từ đó đưa ra gợi ý lựa chọn môn học phù hợp với từng sinh viên cụ thể, nhằm giúp sinh viên và cố vấn học tập giảm nhiều thời gian và công sức, cũng như có được những gợi ý từ hệ thống dự đoán có độ chính xác ngày càng cải thiện.

Việc tích hợp kiến thức học sâu cũng là một đề xuất mới hứa hẹn giúp cải thiện hiệu suất của hệ thống dự đoán và gợi ý. Khi kiến thức học sâu được tích hợp vào hệ thống, nó giúp cho hệ thống có khả năng học và phân tích dữ liệu một cách chính xác và nhanh chóng hơn. Điều này giúp cho hệ thống có thể đưa ra các gợi ý phù hợp hơn với nhu cầu của người dùng và cải thiện độ chính xác của các dự đoán.

Nhìn chung, tích hợp các đề xuất mới này có thể đem lại nhiều lợi ích cho người dùng. Với mối quan hệ xã hội được tích hợp vào hệ thống dự đoán và gợi ý, người dùng có thể dễ dàng tìm kiếm và nhận được các gợi ý liên quan đến những người bạn, người thân của mình. Điều này giúp cho người dùng có thể tiết kiệm được thời gian và tăng cường sự kết nối giữa các thành viên trong mạng lưới của họ.