HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Lê Tuấn Anh

NGHIÊN CỨU PHƯƠNG PHÁP HỌC SÂU CHO HỆ TƯ VẤN

Chuyên ngành: Hệ thống thông tin
Mã số: 8.48.01.04

TÓM TẮT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

HÀ NỘI – 2024

Đề án tốt nghiệp được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: TS. Nguyễn Duy Phương

Phản biện 1: PGS.TS. Phan Xuân Hiếu

Phản biện 2: PGS.TS. Hoàng Xuân Dậu

Đề án tốt nghiệp sẽ được bảo vệ trước Hội đồng chấm đề án tốt nghiệp thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: 09 giờ 15 phút ngày 20 tháng 03 năm 2024

Có thể tìm hiểu đề án tốt nghiệp tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông.

MỞ ĐẦU

Trong những năm gần đây, với sự phổ biến ngày càng lớn của dữ liệu và ứng dụng trực tuyến, hệ tư vấn đối diện với nhiều thách thức. Ngày càng khó khăn để xử lý và phân tích lượng lớn dữ liệu, đồng thời cần tìm ra cách cá nhân hóa đề xuất sao cho phù hợp và chính xác. Đây là một trong những lý do phương pháp học sâu nổi lên như một công cụ mạnh mẽ để giải quyết những thách thức này. Học sâu mang lại khả năng học và rút trích đặc trưng từ dữ liệu lớn, xử lý thông tin phức tạp, và từ đó, cải thiện độ chính xác và khả năng đề xuất của hệ tư vấn.

Hướng nghiên cứu hiện tại của hệ thống tư vấn có thể được chia thành ba loại [5]: khuyến nghị dựa trên nội dung (Content-based Filtering Recommendation), khuyến nghị dựa trên lọc cộng tác (Collaborative Filtering Recommendation) và phương pháp khuyến nghị kết hợp (Hybrid Recommendation). Lọc thông tin theo nội dung khai thác những khía cạnh liên quan đến nội dung thông tin sản phẩm hoặc người dùng đã từng tương tác trong quá khứ để tạo nên tư vấn. Trái lại, lọc thông tin theo cộng tác khai thác những khía cạnh liên quan đến thói quen sở thích của người sử dụng sản phẩm để đưa ra dự đoán và phân bổ các sản phẩm cho người dùng này. Các phương pháp khuyến nghị kết hợp tìm cách đạt được kết quả khuyến nghị tốt nhất bằng cách kết hợp các phương pháp khuyến nghị dựa trên nội dung và các phương pháp khuyến nghị dựa trên lọc cộng tác. Trong bối cảnh của việc cải thiện hệ thống gợi ý hiện nay, sự phát triển của các phương pháp lọc cộng tác (*Collaborative filtering*) đã đem lại một cách tiếp cận mạnh mẽ và linh hoạt hơn trong việc cung cấp gợi ý dựa trên hành vi và sở thích của người dùng. Các phương pháp lọc cộng tác có thể được phân thành hai loại Dựa trên bộ nhớ (Memory-Based) và Dựa trên mô hình (Model-Based) [6]. Lọc cộng tác dựa trên mô hình cho kết quả tốt hơn lọc cộng tác dựa vào bộ nhớ.

Chính vì những ưu điểm đã nêu ra ở trên của các phương pháp, em đã lựa chọn đề tài: "*Nghiên cứu phương pháp học sâu cho hệ tư vấn*", phương pháp sẽ xây dựng hệ tư vấn lọc cộng tác dựa trên mạng nơ-ron lọc cộng tác (Neural Collaborative Filtering). Hy vọng rằng nghiên cứu này sẽ mang lại những hiểu biết sâu hơn và khám phá mới trong lĩnh vực này, tạo ra những đóng góp ý nghĩa cho cộng đồng nghiên cứu và mang lại giá trị thực tế cho người dùng.

Nội dung đề án được trình bày thành ba chương theo cấu trúc sau:

Chương 1: Tổng quan về hệ tư vấn

Trình bày tổng quan về các khái niệm liên quan cơ bản tới hệ tư vấn và đưa ra các phương pháp tiếp cận chính, các ưu và nhược điểm của từng phương pháp này được sử dụng để xây dựng hệ thống tư vấn.

Chương 2: Học sâu cho hệ tư vấn lọc cộng tác

Giới thiệu tổng quan về học sâu, các khái niệm đặc trưng của phương pháp phân rã ma trận, ưu và nhược điểm của phương pháp này. Trình bày chi tiết cách xây dựng mô hình Neural Collaborative Filtering (NCF) để giải quyết bài toán.

Chương 3: Thực nghiệm và kết quả

Xây dựng bộ dữ liệu từ dữ liệu thực tế, trình bày quá trình cài đặt thử nghiệm, so sánh hiệu suất của phương pháp Neural Collaborative Filtering (NCF) với một số phương pháp hiện có.

Kết luận và hướng phát triển.

Trình bày tóm tắt những kết quả đã đạt và chưa đạt được. Qua đó đề xuất mục tiêu, hướng nghiên cứu, cũng như hướng phát triển tiếp theo.

CHƯƠNG 1. TỔNG QUAN VỀ HỆ TƯ VẤN

1.1. Giới thiệu về hệ tư vấn

Trong xã hội ngày nay, vai trò của hệ tư vấn trở nên quan trọng không thể thiếu trong cuộc sống hàng ngày. Nó không chỉ giúp người dùng tiết kiệm thời gian và công sức trong việc tìm kiếm thông tin phù hợp với nhu cầu cá nhân mà còn mang đến trải nghiệm cá nhân hóa, tối ưu và thú vị.

1.1.1. Giới thiệu bài toán tư vấn

Cho một tập hợp hữu hạn $U = \{u_1,\ u_2, \dots, u_n\}$ là tập gồm *N* người dùng (người sử dụng hệ thống), $I = \{i_1,\ i_2, \dots, i_k\}$ là tập gồm *K* sản phẩm (sản phẩm của hệ thống). Mỗi sản phẩm $i_k \in I$ có thể là sản phẩm hàng hóa, tài liệu, sách, báo, hoặc bất kể dạng thông tin nào mà người dùng quan tâm.

Ma trận đánh giá $A = \{a_{ij},\ i = 1, \dots,\ N,\ j = 1, \dots,\ K\}$ dùng để biểu diễn mối quan hệ giữa tập người dùng U và tập sản phẩm *I*. Mỗi giá trị $a_{ij} \in \{0,1,2, \dots, V\}$ thể hiện đánh giá của người dùng $u_i \in U$ đối với sản phẩm $i_j \in I$. Giá trị của a_{ij} có thể thu thập trực tiếp từ ý kiến của người dùng hoặc thu thập một cách gián tiếp thông qua các cơ chế phản hồi của người dùng. Giá trị $a_{ij} = 0$ có thể hiểu rằng người dùng u_i chưa bao giờ biết đến hoặc chưa đánh giá sản phẩm i_j. Nhiệm vụ của hệ thống gợi ý là dựa trên những dữ liệu đã có, đưa ra những gợi ý về sản phẩm $i_j \in I$ mà người dùng $u_i \in U$ có khả năng sẽ quan tâm.

1.1.2. Một số khái niệm chung về hệ thống tư vấn

Hệ thống tư vấn, còn được gọi là Recommender System hoặc Recommendation System [1], là một loại công nghệ thông tin được thiết kế để tự động đề xuất các mục hoặc sản phẩm mà có thể phù hợp và được ưa thích nhất với mỗi người dùng cá nhân. Mục tiêu chính của hệ thống tư vấn

là cung cấp các gợi ý cá nhân hóa, giúp người dùng khám phá và tiêu thụ nội dung mới một cách hiệu quả

Sản phẩm (*Item*) là thuật ngữ chung để chỉ những thứ mà người dùng có thể tương tác trong hệ thống tư vấn. Item có thể là sách, phim, truyện, tin tức...

Trong thực tế, việc thu thập dữ liệu Hồ sơ người dùng thường được sử dụng thông qua hai phương pháp chính là phản hồi ẩn (implicit feedback) và phản hồi tường minh (explicit feedback). Đối với phương pháp phản hồi tường minh (explicit feedback) hệ thống yêu cầu người dùng thực hiện việc xếp hạng (rating) cụ thể cho mỗi sản phẩm để xây dựng Hồ sơ người dùng. Phương pháp này cung cấp dữ liệu người dùng trực tiếp cho hệ thống tư vấn mà không cần các bước biến đổi trung gian, và kết quả tư vấn từ đó được đánh giá là đáng tin cậy hơn [2].

Để khắc phục những mặt hạn chế của việc thu thập dữ liệu hồ sơ người dùng của phương pháp phản hồi tường minh (explicit feedback), phương pháp thu thập phản hồi ẩn (implicit feedback) sử dụng/ghi nhận các dấu vết mà người dùng có thể để lại trên hệ thống như lịch sử truy cập vào website, lịch sử xem hoặc mua sản phẩm, thời gian truy cập trang web, số lần nhấp chuột, và các hoạt động khác tương tự để có thể suy luận các thông tin về sở thích của họ.

Ma trận tương tác Người dùng – Sản phẩm hay còn được gọi là *Utility Matrix* hoặc *User – Item matrix* là một cơ sở dữ liệu mô tả sở thích của mỗi Người dùng (User) với từng Sản phẩm (Item) trong hệ thống. Dữ liệu này có thể được biểu diễn dưới dạng ma trận, trong đó mỗi hàng đại diện cho một người dùng (User), mỗi cột đại diện cho một sản phẩm (Item), và giá trị tại mỗi ô của ma trận thể hiện giá trị đánh giá (rating) của người dùng đó cho sản phẩm tương ứng.

	Sản phẩm 1	Sản phẩm 2	Sản phẩm 3	...	Sản phẩm n
Người dùng 1	1	5	...	3	...
Người dùng 2	...		2	...	4
...	2	1	...	5	...
Người dùng m	3	...	1	...	1

R

Hình 1. 1: Ma trận tương tác Người dùng – Sản phẩm

1.1.3. Các tiêu chí đánh giá hệ tư vấn

1.1.3.1. Phương pháp đánh giá hệ tư vấn

Để có thể đánh giá được độ chính xác của hệ thống tư vấn/khuyến nghị, đầu tiên từ ma trận đánh giá R, chúng ta tiến hành chia tập người dùng U (các hàng trong ma trận đánh giá R) thành hai phần, một phần ký hiệu là U_{train} được sử dụng làm dữ liệu huấn luyện (training), phần còn lại ký hiệu là U_{test} được sử dụng để kiểm tra (testing) sao cho $U_{train} \cup U_{test} = U$ và $U_{train} \cap U_{test} = \emptyset$ Tập dữ liệu huấn luyện U_{train} được sử dụng để xây dựng mô hình theo các thuật toán sử dụng trong hệ tư vấn/khuyến nghị. Tập dữ liệu kiểm tra U_{test} được sử dụng vào quá trình kiểm nghiệm thuật toán tư vấn. Chúng ta có thể biết đến một số cách tiếp cận thưởng được sử dụng để chia tập người dùng U thành 2 phần huấn luyện (U_{train})và kiểm tra (U_{test}) là: Lấy mẫu Bootstrap *(Bootstrap sampling*), Phân chia (*Splitting*), Kiểm thử chéo (*k-fold cross validation*).

1.1.3.2. Độ đo đánh giá độ chính xác của đánh giá dự đoán

Để đánh giá tính chính xác của các giá trị dự đoán từ hệ tư vấn, một trong những phương pháp phổ biến là sử dụng các độ đo dựa trên độ sai số giữa giá trị dự đoán và giá trị thực tế. Điều này giúp đo lường mức độ chính xác của dự đoán và đưa ra cái nhìn tổng quan về hiệu suất của hệ thống. Một

số độ đo phổ biến được sử dụng để đánh giá sai số trong các bài toán phân loại: Sai số trung bình tuyệt đối (Mean Absolute Error - MAE , Độ đo trung bình lỗi lấy căn (Root Mean Square Error - RMSE)

1.1.3.3. Độ đo đánh giá độ chính xác của danh sách sản phẩm tư vấn

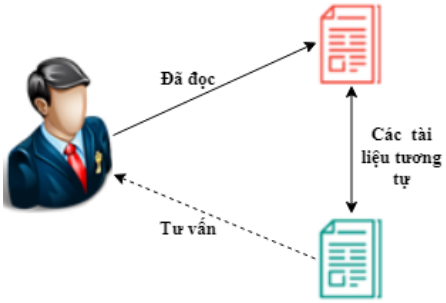
Các độ đo phổ biến được sử dụng để đánh giá độ chính xác của danh sách sản phẩm tư vấn cung cấp một cái nhìn chi tiết về hiệu suất của hệ thống, đặc biệt là trong các hệ thống tư vấn sản phẩm hoặc nội dung. Một số độ đo quan trọng: Precision, Recall, E-measure và F-measure, MAP (Mean Average Precision)

1.2. Các hướng tiếp cận trong hệ tư vấn

Hướng nghiên cứu hiện tại của hệ thống tư vấn có thể được chia thành ba loại [5]: tư vấn dựa trên lọc theo nội dung (*Content-Base Filtering*), tư vấn dựa trên lọc cộng tác (*Collaborative Filtering*) và tư vấn dựa trên lọc kết hợp (*Hybrid*).

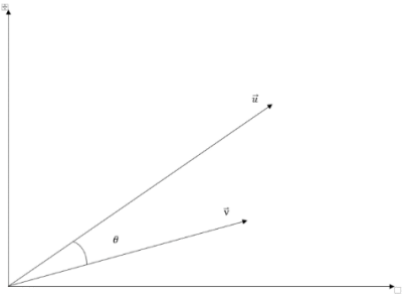
1.2.1. Lọc theo nội dung (Content-Base Filtering)

Phương pháp khuyến nghị dựa trên nội dung *(Content-Base Filtering)* sử dụng nội dung của các sản phẩm và tìm ra điểm tương đồng giữa chúng. Sau khi phân tích đủ số lượng sản phẩm mà một người dùng đã thể hiện sự yêu thích, hồ sơ sở thích của người dùng sẽ được thiết lập.



Hình 1. 2: Cơ chế hoạt động lọc theo nội dung

Hệ thống tư vấn ghi nhận Hồ sơ người dùng (User Profile) dưới dạng vector $\vec{u} = (u_1, u_2, \ldots, u_n)$, trong đó u_i là trọng số thể hiện mức độ quan tâm của người dùng đối với từng thuộc tính của sản phẩm. Vector Hồ sơ sản phẩm (Item Profile) $v = (v_1, v_2, \ldots, v_n)$ biểu diễn thông tin sản phẩm thông qua các thuộc tính v_i. Mức độ phù hợp giữa sở thích của người dùng và sản phẩm được đánh giá dựa trên góc lệch giữa hai vector Hồ sơ người dùng (u) và vector Hồ sơ sản phẩm (v).



Hình 1. 3: Độ tương tự giữa hai vector

Để đánh giá mức độ tương đồng giữa 2 vector $\vec{u}$ và $\vec{v}$ để có thể đưa ra tư vấn, hệ thống thực hiện so sánh bằng cosine góc lệch giữa 2 vector:

$$\text{similarity} = \cos(\theta) = \frac{\vec{u}\vec{v}}{\|\vec{u}\| \cdot \|\vec{v}\|} = \frac{\sum_{i=1}^{n} u_i \cdot v_i}{\sqrt{\sum_{i=1}^{n} u_i^2} \cdot \sqrt{\sum_{i=1}^{n} v_i^2}} \quad (1)$$

1.2.2. Lọc cộng tác (Collaborative Filtering)

Các phương pháp khuyến nghị dựa trên lọc cộng tác *(Collaborative Filtering)* [6] tận dụng tối đa thông tin hành vi và thông tin tùy chọn do người dùng tạo trước đây mà không sử dụng thông tin cá nhân của người dùng và thông tin mô tả sản phẩm, chẳng hạn như đánh giá của người dùng về sản phẩm để tạo sản phẩm được khuyến nghị. Các phương pháp Collaborative

Filtering có thể được phân thành hai loại Dựa trên bộ nhớ (Memory-Based) và Dựa trên mô hình (Model-Based) [7].

Hình 1. 4: Cơ chế hoạt động lọc cộng tác

1.2.2.1. Lọc cộng tác dựa trên bộ nhớ

Dựa trên giá trị đánh giá (rating) của người dùng trong ma trận Người dùng – Sản phẩm (User – Item), độ tương đồng giữa người dùng hiện tại với những người dùng tương tự được tính theo hai bước như sau:

Bước 1: Hệ thống tính toán độ tương tự giữa các người dùng/sản phẩm.

$$\text{similarity}(x,y) = \cos(\vec{x},\vec{y}) = \frac{\vec{x}\cdot\vec{y}}{\|\vec{x}\|_2 \times \|\vec{y}\|_2} = \frac{\Sigma r_{x,i}\cdot r_{y,i}}{\sqrt{\Sigma r_{x,i}^2}\sqrt{\Sigma r_{y,i}^2}} \quad (2)$$

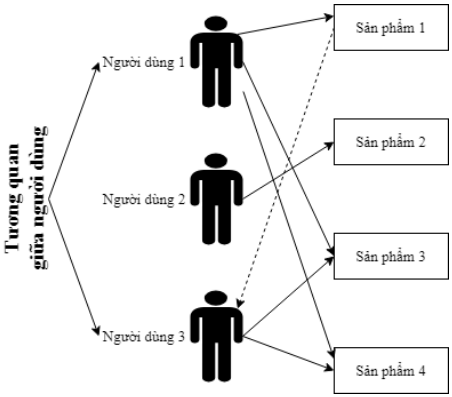
Bước 2: Tính toán giá trị đánh giá dự đoán theo công thức:

$$r(a,i) = \bar{r_a} + \frac{\sum_{i=1}^{n}\left(r_{u,i} - \bar{r_u}\right)\times \text{similarity}(x,y)}{\sum_{i=1}^{n}\text{similarity}(x,y)} \quad (3)$$

Kỹ thuật lọc cộng tác dựa trên bộ nhớ có thể thực hiện dựa trên người dùng và dựa trên đối tượng:

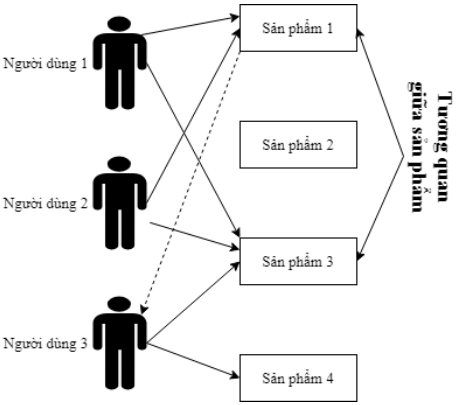
Kỹ thuật dựa trên người dùng: Phương pháp này đo độ tương đồng giữa các người dùng bằng cách so sánh đánh giá (rating) của họ trên các sản

phẩm tương tự. Đánh giá dự đoán cho sản phẩm hiện đang xem xét của người dùng được tính bằng cách lấy giá trị đánh giá (rating) trung bình của sản phẩm đó, nhân với trọng số phản ánh mức độ tương đồng của người dùng hiện tại với những người dùng có sở thích tương tự.



Hình 1. 5: Lọc cộng tác dựa trên bộ nhớ thông qua người dùng

Kỹ thuật lọc dựa trên đối tượng: Phương pháp này tính toán dự đoán đánh giá (rating) dựa trên sự tương đồng giữa các sản phẩm. Nó xây dựng một mô hình các sản phẩm tương tự bằng cách xem xét tất cả các sản phẩm đã được đánh giá bởi người dùng đang hoạt động từ ma trận tương tác. Sau đó, nó xác định mức độ tương đồng của các sản phẩm truy xuất đối với sản phẩm đích, chọn ra k sản phẩm tương tự nhất và tính toán dự đoán bằng cách lấy trung bình có trọng số của đánh giá người dùng trên các sản phẩm tương tự k.



Hình 1. 6: Lọc cộng tác dựa trên bộ nhớ thông qua đối tượng

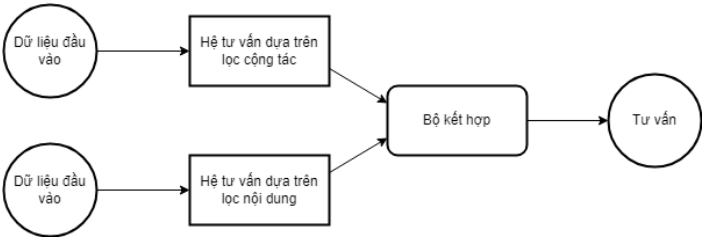
1.2.2.2. Lọc cộng tác dựa trên mô hình

Để giải quyết các hạn chế về thời gian tính toán và yêu cầu dung lượng bộ nhớ lớn, hai tác vụ chính trong phương pháp lọc cộng tác đã được phân biệt rõ ràng. Cụ thể, có hai nhiệm vụ quan trọng: xây dựng mô hình từ dữ liệu huấn luyện và tính toán kết quả tư vấn từ mô hình đã được xây dựng. Để huấn luyện mô hình và tạo ra một mô hình rút gọn mô phỏng sở thích của người dùng, Hệ thống Recommendation System sử dụng các thuật toán học máy, có thể là không giám sát hoặc có giám sát hoặc. Các thuật toán học máy có thể bao gồm cây quyết định, bộ phân loại Bayes, hồi quy, máy vector hỗ trợ (SVM), mạng nơ-ron và các thuật toán khác. Khi đã có mô hình từ quá trình huấn luyện, hệ thống RS sử dụng mô hình này trực tiếp để sản sinh ra kết quả tư vấn.

1.2.3. Lọc kết hợp (Hybrid Filtering)

Các phương pháp khuyến nghị kết hợp *(Hybrid Filtering)* [8] tìm cách đạt được kết quả khuyến nghị tốt nhất bằng cách kết hợp các phương pháp khuyến nghị dựa trên nội dung và các phương pháp khuyến nghị dựa trên lọc cộng tác. Các hệ thống khuyến nghị kết hợp được chia thành khuyến

nghị kết hợp nguyên khối, khuyến nghị kết hợp song song và khuyến nghị kết hợp đường ống [9].



Hình 1. 7: Cơ chế lọc kết hợp

1.2.4. Tư vấn xã hội (Social Recommendation)

Với sự bùng nổ phát triển của Internet, chúng ta có khả năng tiếp cận và khám phá toàn bộ thế giới thông qua những thiết bị thông minh như máy tính và điện thoại di động. Nhu cầu về kết nối và chia sẻ của con người đã điều chỉnh và phát triển theo chiều hướng này. Các bài toán tư vấn xã hội phổ biết là: tư vấn nội dung, tư vấn địa điểm, tư vấn bạn bè

1.3. Phương pháp học sâu trong Collaborative Filtering

Các phương pháp học sâu đã mang lại sự đột phá trong lĩnh vực Collaborative Filtering (CF), mở ra những khả năng mới trong việc cải thiện độ chính xác của các hệ thống đề xuất. Thay vì chỉ tập trung vào sự tương tác giữa người dùng và sản phẩm, các phương pháp này tích hợp sức mạnh của mô hình học sâu để hiểu biểu diễn phức tạp của dữ liệu.

1.4. Kết luận chương

Nội dung chương 1 đã trình bày làm rõ một số khái niệm cơ bản của hệ tư vấn, các phương pháp tiếp cận để xây dựng hệ tư vấn. Chương 1 cũng đưa ra chi tiết những ưu điểm cũng như nhược điểm của từng phương pháp tiếp cận. Bên cạnh đó, chương 1 đã trình khái quát về phương pháp học sâu trong lọc cộng tác làm cơ sở để lựa chọn phương pháp học sâu cho hệ tư vấn được trình bày ở chương 2.

CHƯƠNG 2. HỌC SÂU CHO HỆ TƯ VẤN LỌC CỘNG TÁC

2.1. Giới thiệu về học sâu

Deep Learning (DL) không chỉ là một phần của học máy, mà còn là một lĩnh vực quan trọng và đầy tiềm năng trong lĩnh vực trí tuệ nhân tạo (AI).

Học sâu cũng đóng một vai trò rất quan trọng trong các hệ thống tư vấn/khuyến nghị, nó giúp người dùng có thể có những trải nghiệm tốt nhất. Học sâu không chỉ là một công nghệ mạnh mẽ, mà còn là một lĩnh vực nghiên cứu đầy triển vọng, với tiềm năng để tạo ra những tiến bộ đáng kể trong nhiều lĩnh vực khác nhau trong tương lai.

2.1.1. Cách thức hoạt động của học sâu

Học sâu (Deep Learning - DL) hoạt động bằng cách khám phá và tìm hiểu các cấu trúc phức tạp trong dữ liệu.

Các nơ ron được nhóm vào 3 loại layer khác nhau: Input layer, các hidden layer, Output layer

- *Input layer:* Nhận các dữ liệu đầu vào
- *Các hidden layer*: Thực hiện các phép tính toán trên các đầu vào. Một trong những thách thức lớn khi thiết kế mạng nơ-ron là quyết định về số lượng hidden layer và số nơ-ron trong mỗi layer.
- *Output layer*: trả về dữ liệu đầu ra

2.1.2. Ưu điểm của học sâu so với phương pháp học máy cổ điển

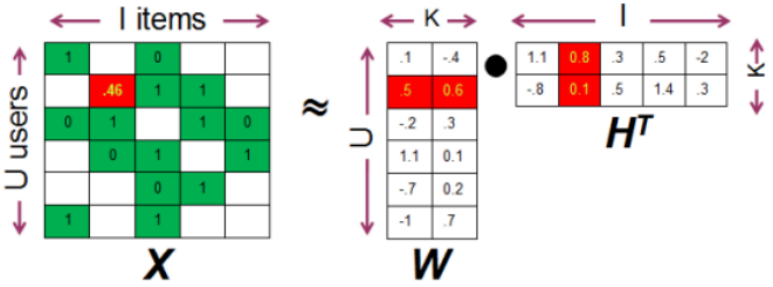
Học sâu mang lại nhiều ưu điểm so với phương pháp machine learning cổ điển, và điều này đã đóng vai trò quan trọng trong việc biến đổi cách chúng ta tiếp cận và giải quyết các bài toán máy học.

2.2. Phương pháp phân rã ma trận

Phân rã ma trận (Matrix Factorization - MF) [12] là một trong những thuật toán lâu đời sử dụng trong các hệ thống gợi ý. Mô hình đầu tiên được

đề xuất bởi Simon Funk và phát triển mạnh mẽ, phổ biến sau cuộc thi Netflix tổ chức năm 2006.

Kỹ thuật phân rã ma trận (Matrix Factorization) được thực hiện thông qua việc chia một ma trận lớn X thành 2 ma trận có kích thước nhỏ hơn rất nhiều so với ma trận ban đầu là W và H, sao cho X có thể được xây dựng lại từ hai ma trận nhỏ hơn này càng chính xác càng tốt [11], điều này có nghĩa là $\mathbf{X} \approx \mathbf{WH^T}$ như minh hoạ trong hình 2.3. Trong đó, X là tập hợp tất cả các đánh giá (rating) của người dùng (user) với mục tin (item), bao gồm cả những giá trị chưa biết cần được dự đoán tạo nên một ma trận gọi là Utility Matrix.



Hình 2. 1: Minh họa kỹ thuật phân rã ma trận

Trong đó:

- K là số nhân tố tiềm ẩn, và nhỏ hơn rất nhiều so với số người dùng và số mục dữ liệu (K << $|U|$ và K << $|I|$).
- W ∈ $R^{|U|\times K}$ là một ma trận mà mỗi dòng là một vector bao gồm K nhân tố tiềm ẩn (latent factors) mô tả user u.
- W ∈ $R^{|I|\times K}$ là một ma trận mà mỗi dòng là một vectors bao gồm K nhân tố tiềm ẩn mô tả cho item i.

Ý tưởng chính của MF là tồn tại các tính chất ẩn (latent features) mô tả sự liên quan giữa các người dùng (user) và sản phẩm (item). Mỗi người dùng sẽ mang chất ẩn nào đó và được mô tả bởi các hệ số trong vector w_u.

Tương tự, mỗi sản phẩm cũng sẽ có những tính chất ẩn ở một mức độ nào đó tương ứng với các hệ số trong vector h_i. Hệ số càng cao tương ứng với việc mang tính chất đó càng cao.

Gọi h_{ik}và w_{uk}là các phần tử tương ứng của hai ma trận H và W, khi đó xếp hạng (rating) của người dùng u trên mục dữ liệu i được dự đoán bởi công thức (4):

$$\hat{r}_{ui} = \sum_{k=1}^{K} w_{uk} h_{ik} = (WH^T)_{u,i} \quad (4)$$

2.3. Phương pháp học sâu cho hệ tư vấn

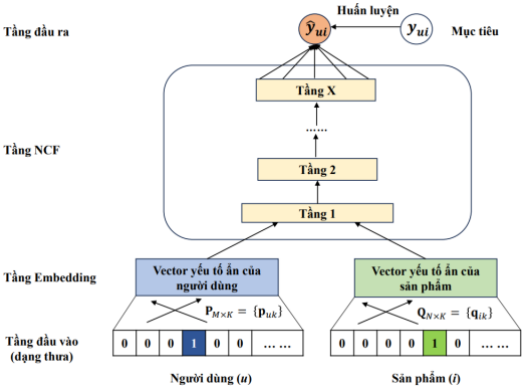
Do phương pháp Matrix Factorization (MF) còn có nhiều hạn chế như đã nêu ở mục 2.2 nên việc phải thiết kế một hàm tương tác tốt hơn để mô hình hóa việc tương tác giữa các thuộc tính ẩn (latent features) của người dùng (user) và sản phẩm (item) là vô cùng cần thiết. Vì vậy đề án đề xuất phương pháp học sâu Neural Collaborative Filtering (NCF) cho hệ tư vấn. Mô hình NCF đi khai thác phương pháp MF theo hướng này và thực hiện nó bằng mạng nơ-ron.

2.3.1. Phương pháp Neural Collaborative Filtering

NCF kết hợp cả hai tiếp cận truyền thống trong việc xây dựng hệ thống đề xuất: Collaborative Filtering (CF) và Neural Networks (NN), từ đó tạo ra một phương pháp mạnh mẽ và linh hoạt. Ở bản chất, CF tập trung vào việc dự đoán sở thích của một người dùng dựa trên lịch sử hoặc hành vi của người dùng đó cũng như hành vi của những người dùng khác. Trong khi đó, NN có khả năng học các biểu diễn phức tạp từ dữ liệu đầu vào và thực hiện các dự đoán dựa trên các biểu diễn đó.

NCF kết hợp cả hai mô hình này bằng cách sử dụng một mạng nơ-ron để học các biểu diễn người dùng và mặt hàng từ dữ liệu, sau đó sử dụng các biểu diễn này để dự đoán sở thích hoặc xếp hạng của người dùng cho các

mặt hàng mà họ chưa xem hoặc đánh giá. Mô hình NCF được nhóm tác giả Xiangnan He và cộng sự [13] đưa ra vào năm 2017. Nhóm tác giả đã đưa ra đề xuất sử dụng cấu trúc NCF tổng quát được thể hiện như trong hình vẽ dưới đây:



Hình 2. 2: Kiến trúc mô hình NCF

Trong hình vẽ trên, đi từ dưới lên, mô hình NCF bao gồm các tầng: (1) Tầng đầu vào, (2) Tầng Embedding, (3) Tầng NCF, (4) Tầng đầu ra.

2.3.1.1. Tầng đầu vào

Tầng đầu vào bao gồm các thuộc tính vector của *người dùng* và *sản phẩm*. Như trong mô hình trên, các vector đó là biểu diễn dạng one-hot là vector với chỉ một trường có giá trị bằng 1, các trường còn lại có giá trị 0 dùng để biểu diễn định danh của *người dùng* (userID) và *sản phẩm* (itemID). Đây là biểu diễn thường thấy của các phương pháp cổ điển.

2.3.1.2. Tầng Embedding

Dữ liệu phân loại (categorical data) là các loại dữ liệu mà các thuộc tính của chúng nhận giá trị từ một tập hữu hạn các danh mục khác nhau, thay vì nhận các giá trị liên tục. Dữ liệu này sẽ được biểu diễn dưới dạng tensor một chiều hoặc vector $\vec{V}$ có các phần tử nhận giá trị như sau:

$$v_j = \begin{cases} 1 & \text{nếu nhận giá trị} \\ 0 & \text{nếu không nhận giá trị} \end{cases} \quad (7)$$

Thuật toán mã hóa được sử dụng ở đây được gọi là one-hot encoding, trong đó đầu ra của thuật toán là vector $\vec{V}$, một vector thưa có ít phần tử khác 0. Trong ngữ cảnh của hệ thống tư vấn, ma trận Người dùng – Sản phẩm (User – Item) có thể được hiểu như một tập hợp các tensor người dùng, trong đó các phần tử tương ứng với các sản phẩm mà người dùng đã đánh giá (rating).

Tuy nhiên người dùng thường chỉ xem một phần nhỏ các bộ phim trong toàn bộ tập hợp các bộ phim. Có thể những vấn đề nêu trên là bằng cách sử dụng các thuật toán nhúng (embedding), trong đó vector thưa có số chiều lớn được ánh xạ vào một không gian vector có số chiều thấp hơn dày đặc hơn. Quá trình nhúng này giúp tạo ra các vector đặc tính ẩn (latent vector) mô tả về đối tượng (sản phẩm hoặc người dùng) một cách dày đặc và bảo toàn các mối quan hệ ngữ nghĩa tồn tại trong dữ liệu ban đầu.

Lớp embedding nhận 03 tham số đầu vào quan trọng như sau:

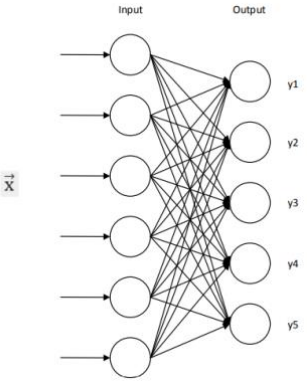
- data: mảng số nguyên chứa Id của đối tượng được embedding.
- input_dim: độ lớn của mảng data.
- output_dim: kích thước của vector đặc tính ẩn.

Các đặc tính ẩn thu được từ lớp embedding có thể là những đặc điểm về nội dung của bộ phim mà người dùng ưa thích, như thể loại phim, năm sản xuất, mức độ lời thoại, và nhiều yếu tố khác.

2.3.1.3. Tầng Neural Collaborative Filtering

Các tầng thuộc mạng nơ-ron (tầng 1 đến N) - Tầng neural collaborative: các tầng này có nhiệm vụ ánh xạ từ các vector embedding đến vector đầu vào của tầng đầu ra (output). Mỗi tầng là một lớp Multi-layer Perceptron (MLP).

Lớp ẩn trong mạng nơ-ron được sử dụng để mô hình hóa các quan hệ phi tuyến tính có thể tồn tại trong không gian vector đặc tính ẩn nằm ở đầu ra của lớp nối



Hình 2. 3: Hoạt động của lớp ẩn

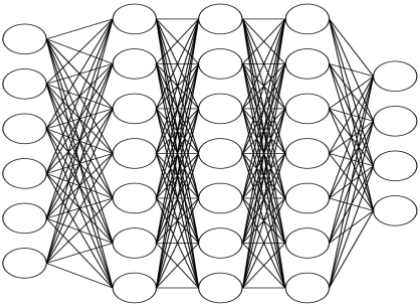
Cho $x \in \mathbb{R}^m$ $y_i \in \mathbb{R}$

$$y_i = \delta(w_1 x_1 + \ \dots \ + w_m x_m) \quad (8)$$

Vector y đầy đủ như sau:

$$y = \begin{pmatrix} \sigma(w_{1,1}x_1 + \dots + w_{1,m}x_m) \\ \vdots \\ \sigma(w_{n,1}x_1 + \dots + w_{n,m}x_m) \end{pmatrix} \quad (9)$$

Khi xếp chồng nhiều lớp mạng ẩn lên nhau, ta thu được mạng Multi-Level Perceptrons MLP như hình 2.8:



Hình 2. 4: Kiến trúc MLP

Đầu ra tại mỗi lớp bằng tích của vector đầu vào với ma trận trọng số và mỗi phần tử của vector đầu ra được áp dụng toán tử phi tuyến σ:

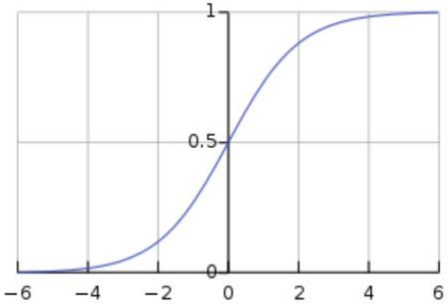
$$y = \sigma(wx) \quad (10)$$

Trong đó: σ là ma trận không gian $R^{m \times n}$ có các phần tử được áp dụng toán tử phi tuyến σ.

Toán tử phi tuyến σ của các lớp là hàm sigmoidal được định nghĩa như sau:

$$\delta(x) = \frac{1}{1 + e^{-x}} \quad (11)$$

Hàm sigmoid được sử dụng để chuyển đổi đầu vào từ một số thực thành một giá trị nằm trong khoảng (0, 1). Đối với đầu vào là số thực dương với giá trị rất lớn sẽ cho đầu ra tiệm cận 1, đầu vào là số thực âm với giá trị rất nhỏ sẽ cho đầu ra tiệm cận 0.



Hình 2. 5: Hàm Sigmoid

Một trong những đặc điểm nổi bật của các mạng ẩn là khả năng học và ghi nhớ thông tin từ dữ liệu huấn luyện theo thời gian. Điều này thường được thể hiện qua việc giảm dần sai số trên tập huấn luyện sau mỗi lượt huấn luyện (epoch).

Sau khi đi qua tất cả các lớp ẩn (Fully-connected layers), vector ẩn đầu ra tại lớp cuối cùng được tính toán theo công thức truy hồi như sau:

$$z_1 = \Phi_1(x),$$

$$\Phi_2(x_1) = \delta_2(W_2^T x_1 + b_2), \qquad (12)$$
$$\dots$$
$$\Phi_L(x_{L-1}) = \delta_L(W_L^T x_{L-1} + b_L),$$

Trong đó:

W_x : Ma trận trọng số tương ứng với lớp thứ x.

b_x: Vector hệ số thiên lệch (bias) tương ứng lớp thứ x.

σ: Hàm kích hoạt của perceptron thứ x.

Đầu ra của mạng MLP là một vector đặc tính ẩn mà trong đó mỗi phần tử mô tả một thuộc tính nào đó của người dùng, sản phẩm, thông tin bổ trợ.

2.3.1.4. Tầng đầu ra

Ở tầng này, điểm số dự đoán $\hat{y}_{ui}$ được tính toán và sử dụng lan truyền ngược để cập nhật các tham số sao cho giá trị $\hat{y}_{ui}$ đạt xấp xỉ y_{ui}. Việc huấn luyện NCF được thực hiện bằng cách tối thiểu hóa hàm lỗi giữa điểm số dự đoán với giá trị mục tiêu (target) y_{ui} tương ứng.

Điểm số dự đoán được tính như sau:

$$\hat{y}_{ui} = f\left(\mathbf{P}^{\mathrm{T}}\mathbf{v}_{\mathrm{u}}^{\mathrm{U}}, \mathbf{Q}^{\mathrm{T}}\mathbf{v}_{\mathrm{i}}^{\mathrm{I}} | \mathbf{P}, \mathbf{Q}, \boldsymbol{\theta}_{\mathrm{f}}\right) \qquad (13)$$

Trong đó:

- v_u^U, v_i^U lần lượt là các ma trận hàng (vector) one-hot biểu diễn cho user u và item i.
- $P \in R^{M \times K}$ và $Q \in R^{N \times K}$ lần lượt là các ma trận tầng embedding chiếu v_u^U, v_i^U vào không gian các thuộc tính ẩn. **P** và **Q** là ma trận trọng số giữa tầng input và tầng embedding.
- Θ_f là tập các tham số mô hình của hàm f. Do hàm f ở đây là ánh xạ đầu vào đầu ra của một mạng nơ-ron nên ta có thể viết f theo công thức sau:

$$f\left(\mathbf{P}^{\mathrm{T}}\mathbf{v}_{\mathrm{u}}^{\mathrm{U}}, \mathbf{Q}^{\mathrm{T}}\mathbf{v}_{\mathrm{i}}^{\mathrm{I}}\right)$$
$$= \emptyset_{out}\left(\emptyset_X\left(\ldots\emptyset_2\left(\emptyset_1(\boldsymbol{P}^T\boldsymbol{v}_u^U, \boldsymbol{Q}^T\boldsymbol{v}_i^I)\right)..\right)\right) \quad (14)$$

Trong đó:

- ϕ_{out} là hàm ánh xạ đầu vào tới đầu ra của tầng output.
- $\phi_n, \phi_{n-1}, \ldots, \phi_2, \phi_1$ lần lượt là các hàm ánh xạ đầu vào tới đầu ra tại các tầng n, n-1, . . ., 2, 1 trong mạng nơ-ron.

2.4. Kết luận chương

Nội dung chương 2 đã trình bày khái niệm của phương pháp phân ra ma trận, ưu và nhược điểm của phương pháp này, lựa chọn phương pháp học sâu cho hệ tư vấn là phương pháp Neural Collaborative Filtering và trình bày kiến trúc mô hình của phương pháp này. Ở chương 3 tiếp theo, đề án sẽ trình bày về tập dữ liệu và tiến hành thực nghiệm và đánh giá kết quả phương pháp học sâu Neural Collaborative Filtering (NCF) đã lựa chọn.

CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ

3.1. Tập dữ liệu thực nghiệm

3.1.1. Tập dữ liệu MovieLens-1M

Movielens [14] là một nhóm nghiên cứu cung cấp các bộ dữ liệu cho các bài toán xây dựng hệ thống gợi ý. Các bộ dữ liệu trong tập này bao gồm thông tin đánh giá xếp hạng của người dùng tới các bộ phim. Những thông tin về người dùng hay các bộ phim cũng được cung cấp. Bộ dữ liệu MovieLens-1M bao gồm xấp xỉ 1 triệu bộ (user, movie, rating) từ khoảng 3900 bộ phim và 6040 người dùng. Số lượng bộ phim mà người dùng đánh giá nhiều nhất là 2314 bộ phim và 20 là số lượng bộ phim ít nhất mà mỗi người dùng đưa ra.

Phân phối của các điểm đánh giá (rating) cho thấy rằng rằng điểm 4 được đánh giá nhiều nhất trong khi các điểm 1 và 2 có ít lượng đánh giá nhất. Điều này có thể được giải thích bằng sự thật là người dùng thường đánh giá khi họ rất thích một bộ phim; khi họ không thực sự thích, họ sẽ ít ra đánh giá hơn.

3.1.2. Tập dữ liệu Dlab

Dlab [15] là cổng lập trình trực tuyến của Khoa Công nghệ thông tin 1 - Học viện Công nghệ Bưu chính Viễn thông. Dữ liệu Dlab được thu thập trực tiếp từ cổng lập trình. Tập dữ liệu thu thập từ 6120 người dùng. Trong đó, người dùng i lập trình đúng bài toán x cổng lập trình ghi lại giá trị 1, người dùng i lập trình chưa đúng bài toán x cổng lập trình ghi lại giá trị -1, người dùng i chưa giải bài toán x cổng lập trình ghi lại giá trị 0. Mỗi người lập trình có thể submit một bài nhiều lần và hệ thống chỉ ghi nhận giá trị 1, -1 cho kết quả cuối cùng. Trong số 6120 người dùng đã lọc ra được 5311 người dùng đã tham gia lập trình ít nhất 20 bài dù đúng hoặc sai để tiến hành thử nghiệm.

3.2. Phương pháp thực nghiệm và kết quả

3.2.1. Phương pháp thực nghiệm

Đầu tiên, toàn bộ dữ liệu thực nghiệm được chia thành hai phần, một phần được sử dụng làm dữ liệu huấn luyện (training) ký hiệu là:U_{tr}, phần còn lại được sử dụng để kiểm tra (testing) ký hiệu là: U_{te}. Tập Utr chứa 80% đánh giá và tập U_{te} chứa 20% đánh giá. Dữ liệu được sử dụng để thực nghiệm và so sánh đánh giá phương pháp Neural Collaborative Filtering (NCF) so với các phương pháp Collaborative Filtering (CF) và Matrix Factorization (MF).

3.2.2. Độ đo đánh giá

Mean Squared Error (MSE) là sai số bình phương trung bình, tính bình phương của sai số giữa giá trị dự đoán và giá trị thực tế, sau đó lấy trung bình của các bình phương sai số đó, theo công thức dưới đây:

$$\mathrm{MSE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{r_i} - \hat{\mathrm{r}}_\mathrm{i})^2 \quad (15)$$

Trong đó:

- N là số điểm dữ liệu
- r_i là giá trị thực
- $\hat{r}_i$ là giá trị dự đoán.

MSE càng thấp thì dự báo càng tốt.

3.2.3. Kết quả thực nghiệm

Tiến hành thực nghiệm phương pháp Neural Collaborative Filtering (NCF) so với các phương pháp Collaborative Filtering (CF) và Matrix Factorization (MF). Các tham số được cài đặt ở mô hình NCF với các tập dữ liệu MovieLens-1M và Dlab cụ thể như sau:

Bộ dữ liệu MovieLens-1M:

- Số thuộc tính ẩn: 150
- Số lớp ẩn: 3

- Số nút ở mỗi lớp ẩn: 250
- Tỷ lệ dropout được áp dụng sau mỗi lớp ẩn: 0.2

Bộ dữ liệu Dlab

- Số thuộc tính ẩn: 64
- Số lớp ẩn: 3
- Số nút ở mỗi lớp ẩn: 150
- Tỷ lệ dropout được áp dụng sau mỗi lớp ẩn: 0.2

Các tập dữ liệu được chia ngẫu nhiên với mỗi lần thực nghiệm với dữ liệu huấn luyện chiếm 80%, dữ liệu kiểm tra chiếm 20%. Đề án sử dụng độ đo MSE để đánh giá và thu được kết quả ở bảng 3.1, kết quả này là kết quả tốt nhất của các lần thực nghiệm.

	MovieLens-1M	**Dlab**
Collaborative Filtering (CF)	0.9313	0.0742
Matrix Factorization (MF)	0.7870	0.0547
Neural Collaborative Filtering (NCF)	**0.7685**	**0.0515**

Bảng 3. 1: Kết quả thực nghiệm

Giá trị MSE trong bảng 3.1 cho thấy phương pháp đề xuất Neural Collaborative Filtering (NCF) cho giá trị MAE nhỏ hơn phương pháp Collaborative Filtering (CF) và Matrix Factorization (MF) trên hai tập dữ thực nghiệm là Movielen-1M và Dlab. Cụ thể, trong trường hợp dữ liệu Movielen-1M có giá trị đánh giá (rating) từ 0.5 đến 5 thì giá trị MSE của phương pháp CF, MF, NCF lần lượt là 0.9313, 0.7870, 0.7685. Giá trị MSE của các phương pháp CF, MF, NCF trên tập dữ liệu Dlab lần lượt là 0.0742,

0.0547, 0.0515. Kết quả thu được pháp Neural Collaborative Filtering (NCF) có giá trị MAE nhỏ hơn các phương pháp còn lại.

3.3. Kết luận chương

Chương 3 đã trình bày về dữ liệu và quá trình thực nghiệm trên tập dữ liệu thực tế áp dụng phương pháp đã đề xuất ở chương 2 để thực hiện so sánh và đánh giá. Kết quả thu được phương pháp đề xuất đạt kết quả tốt và có thể hỗ trợ cho việc đưa tư vấn.

KẾT LUẬN VÀ KIẾN NGHỊ

Đề án này tập trung nghiên cứu về phương pháp học sâu cho hệ tư vấn. Cụ thể đề án đã đạt được các kết quả sau:

- Nghiên cứu tổng quan về hệ tư vấn, các khái niệm và các phương pháp tiếp cận trong hệ tư vấn
- Nghiên cứu về phương pháp học sâu trong hệ tư vấn.
- Đưa ra phương pháp học sâu cho hệ tư vấn và tiến hành xây dựng dữ liệu và thực nghiệm so sánh đánh giá phương pháp đề xuất với các phương pháp khác trên tập dữ liệu xây dựng được.

Do thời gian thực hiện đề án không nhiều nên tác giả chưa có điều kiện nghiên cứu thêm nhiều phương pháp. Trong tương lai, nếu có điều kiện, tác giả sẽ tập trung nghiên cứu để xây dựng, cải tiến các phương pháp học sâu và tiến hành thực nghiệm thêm trên các tập dữ liệu thực tế khác để áp dụng vào các hệ tư vấn.