THIẾT KẾ KHUNG AUTOREC TỔNG QUÁT CHO HỆ THỐNG GỢI Ý

Phạm Thị Thu Trang[1]
Email: ptttrang@hou.edu.vn

Ngày tòa soạn nhận được bài báo: 02/06/2025
Ngày phản biện đánh giá: 15/08/2025
Ngày bài báo được duyệt đăng: 29/08/2025

DOI: 10.59266/houjs.2025.709

Tóm tắt: *Trong những năm gần đây, các hệ thống gợi ý dựa trên lọc cộng tác sử dụng mạng nơ-ron đã đạt được nhiều bước tiến đáng kể trong việc cung cấp các đề xuất cá nhân hóa cho người dùng. Mô hình CI-Autorec cho phép sử dụng thông tin dựa trên nội dung, góp phần nâng cao hiệu quả của hệ thống. Trong khi đó, Sparse Autoencoder đã chứng minh được khả năng tổng hợp và giảm thiểu chi phí tính toán mà vẫn đảm bảo chất lượng kết quả. Bài báo này đề xuất một mô hình kết hợp, tận dụng khả năng khai thác thông tin nội dung của CI-Autorec và sử dụng các đơn vị kích hoạt của Sparse Autoencoder để dự đoán đánh giá của người dùng. Kết quả thực nghiệm cho thấy hệ thống gợi ý đề xuất vượt trội hơn so với các mô hình kết hợp hiện có cả về độ chính xác lẫn hiệu suất.*

Từ khóa: *hệ thống gợi ý, học sâu, mạng tự mã hóa, thông tin dựa trên nội dung, dữ liệu thưa*

I. Đặt vấn đề

Trong những thập kỷ gần đây, sự bùng nổ của mạng Internet đã góp phần thúc đẩy mạnh mẽ quá trình phát triển và chia sẻ thông tin. Cùng với đó, người dùng ngày càng bị bao quanh bởi vô số lựa chọn đến từ các nền tảng mạng xã hội, quảng cáo số và đặc biệt là các trang thương mại điện tử. Điều này đặt ra yêu cầu cấp thiết đối với các nhà cung cấp dịch vụ trong việc giữ chân người dùng và nâng cao tính cạnh tranh thông qua các giải pháp gợi ý cá nhân hóa. Do đó, các hệ thống gợi ý (Recommendation Systems - RS) ngày càng đóng vai trò quan trọng, khi giúp người dùng dễ dàng tiếp cận các nội dung phù hợp với sở thích cá nhân thông qua những đề xuất tự động.

Trong các hệ thống lọc cộng tác (Collaborative Filtering - CF), người dùng có sở thích tương đồng được xác định dựa trên lịch sử tương tác, từ đó đưa ra các đề xuất mà không cần đến thông tin chi tiết về sản phẩm hoặc dịch vụ (Su and Khoshgoftaar, 2009; Duong et al., 2023). Tuy nhiên, độ chính xác của các hệ thống lọc cộng tác dựa trên hàng xóm thường bị ảnh hưởng khi dữ liệu bị thiếu hoặc thưa,

[1] Trường Đại học Mở Hà Nội

gây khó khăn trong việc tìm ra các người dùng hay mặt hàng tương tự. Một số phương pháp bổ sung giá trị (imputation) đã được đề xuất nhằm giảm thiểu sự thưa thớt trong ma trận tương tác, tuy nhiên mức độ đáng tin cậy của các giá trị bổ sung vẫn là vấn đề cần được nghiên cứu thêm (Ranjbar et al., 2015; Duong et al., 2022; Duong et al., 2018). Ngoài ra, các mô hình CF còn gặp phải vấn đề "khởi đầu lạnh" (cold-start), đặc biệt trong trường hợp người dùng mới hoặc có rất ít đánh giá.

Để khắc phục những hạn chế trên, các phương pháp lai (hybrid) đã được nghiên cứu và triển khai. Một ví dụ tiêu biểu là mô hình Factorization Machine (FM), kết hợp kỹ thuật phân rã ma trận với các phương pháp học có giám sát như Máy Vector Hỗ trợ (SVM), cho phép sử dụng đồng thời phản hồi rõ ràng (explicit feedback) và dữ liệu bổ sung (Idrissi and Zellou, 2020; Rendle, 2010). Các mô hình này cho thấy hiệu quả vượt trội, đặc biệt trong bối cảnh dữ liệu có tính thưa cao. Bên cạnh đó, một số nghiên cứu gần đây đã khai thác thêm các thông tin bên ngoài nhằm cải thiện chỉ số tương đồng, qua đó giảm thiểu tác động của vấn đề dữ liệu thưa (Guo et al., 2014; Niu et al., 2016).

Trong bối cảnh đó, mạng nơ-ron được chú ý nhờ khả năng xấp xỉ mọi hàm liên tục (Hornik et al., 1989; Hoa et al., 2023; Son et al., 2024), từ đó trở thành công cụ hữu hiệu trong việc nâng cao khả năng biểu diễn của mô hình (He et al., 2017; He and Chua, 2017). Trong số các mô hình mạng nơ-ron, Autoencoder (AE) nổi bật với khả năng tự động học biểu diễn đặc trưng bằng cách mã hóa và tái tạo dữ liệu đầu vào (Hinton and Salakhutdinov, 2006). Lớp ẩn của AE chứa đựng các đặc trưng quan trọng của dữ liệu đầu vào và được sử dụng để học biểu diễn. Mô hình Autorec, một dạng CF dựa trên AE, đã giới thiệu các phép biến đổi phi tuyến trong CF, dựa trên phản hồi rõ ràng (Sedhain et al., 2015; Zhang et al., 2017).

Trong nghiên cứu trước đây, mô hình CI-Autorec (Duong et al., 2023) được đề xuất với khả năng khai thác thông tin dựa trên nội dung bằng cách đưa thêm các vector thông tin phụ vào đầu vào của encoder, từ đó nâng cao hiệu quả trích xuất đặc trưng và tái tạo đầu ra. Song song với đó, Sparse Autoencoder - một biến thể của AE - đã chứng minh hiệu quả trong việc học không giám sát bằng cách áp dụng ràng buộc thưa (sparsity constraint) lên các đơn vị ẩn (Ng et al., 2011). Điều này giúp mã hóa dữ liệu có chiều cao một cách hiệu quả, trích xuất các đặc trưng có ý nghĩa, đồng thời làm giảm số lượng đơn vị kích hoạt không cần thiết.

Xuất phát từ mong muốn tận dụng đồng thời ưu điểm của hai mô hình CI-Autorec và Sparse Autoencoder, nghiên cứu này đề xuất một kiến trúc kết hợp mới. Mô hình đề xuất không chỉ phát huy thế mạnh riêng biệt của từng kỹ thuật, mà còn khai thác hiệu quả tính bổ sung lẫn nhau giữa chúng, nhằm cải thiện hiệu quả gợi ý cũng như khả năng diễn giải của hệ thống.

Phần còn lại của bài báo được tổ chức như sau: Phần II trình bày các nghiên cứu liên quan. Phần III mô tả chi tiết kiến trúc mô hình được đề xuất. Phần IV phân

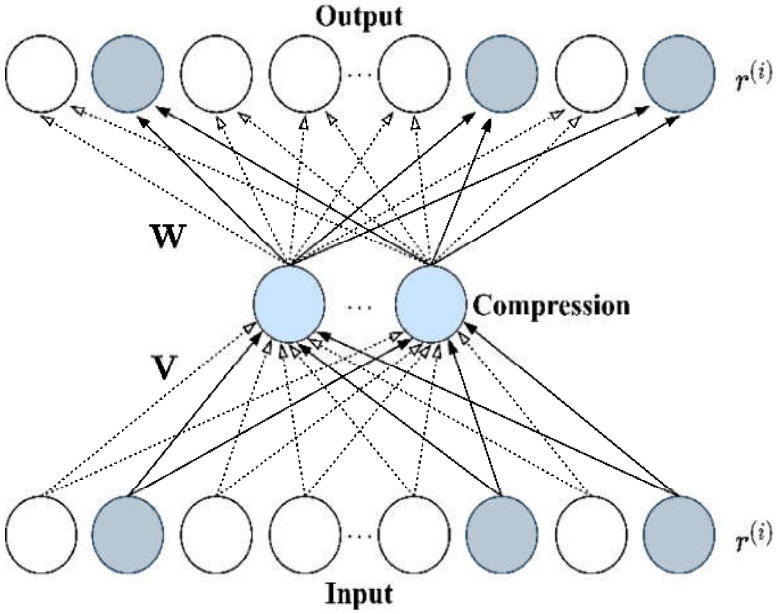
tích kết quả thực nghiệm và so sánh với các mô hình hiện tại. Cuối cùng, Phần V đưa ra kết luận và hướng nghiên cứu trong tương lai.

II. Các nghiên cứu liên quan

2.1. Mô hình Autorec

Trong các phương pháp được sử dụng phổ biến trong truy xuất thông tin và khai phá dữ liệu cho hệ thống gợi ý, Autoencoder (AE) được đánh giá là một trong những cơ chế học biểu diễn hiệu quả nhất (Hinton and Salakhutdinov, 2006). Nhiều nghiên cứu đã ứng dụng thành công kiến trúc AE và các biến thể của nó vào các hệ thống gợi ý (Sedhain et al., 2015; Zhang et al., 2017; Ouyang et al., 2014; Duong et al., 2020)

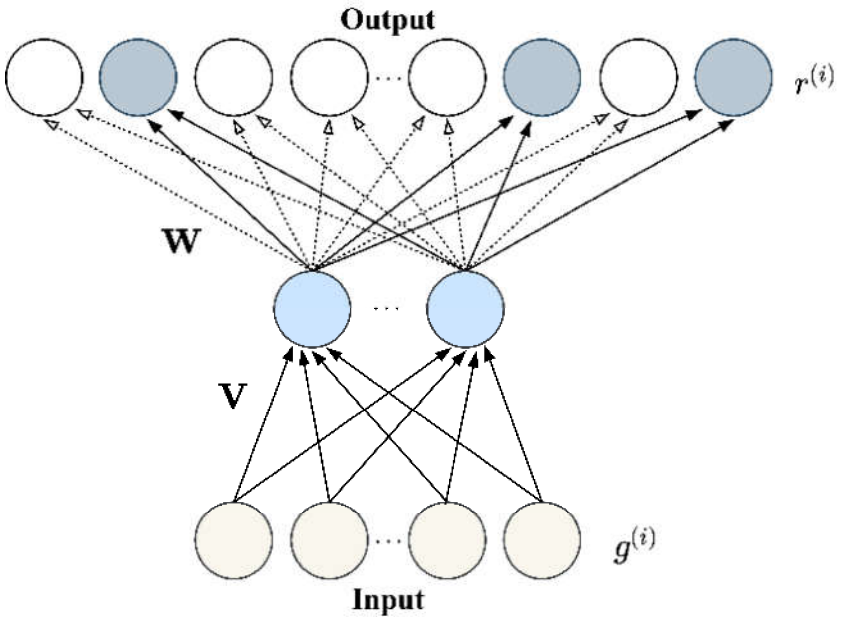
Trong đó, Autorec dựa trên kiến trúc AE thuần túy, nổi bật là mô hình Autorec hướng sản phẩm (I-Autorec) (Sedhain et al., 2015), cho phép tái tạo lại các vector đánh giá đầy đủ từ đầu vào rời rạc, vốn chỉ chứa các đánh giá đã biết. Khác với AE truyền thống, I-Autorec chỉ cập nhật trọng số liên quan đến các đánh giá đã được quan sát nhằm giảm sai số trong dự đoán (Hình 1). Nhờ vậy, kiến trúc này vừa đơn giản trong triển khai, vừa giữ được ưu điểm phi tuyến của AE so với các mô hình nhân tố tiềm ẩn truyền thống.



Hình 1: Kiến trúc mạng I-Autorec.

2.2. Mạng Autorec dựa trên nội dung CI-Autorec

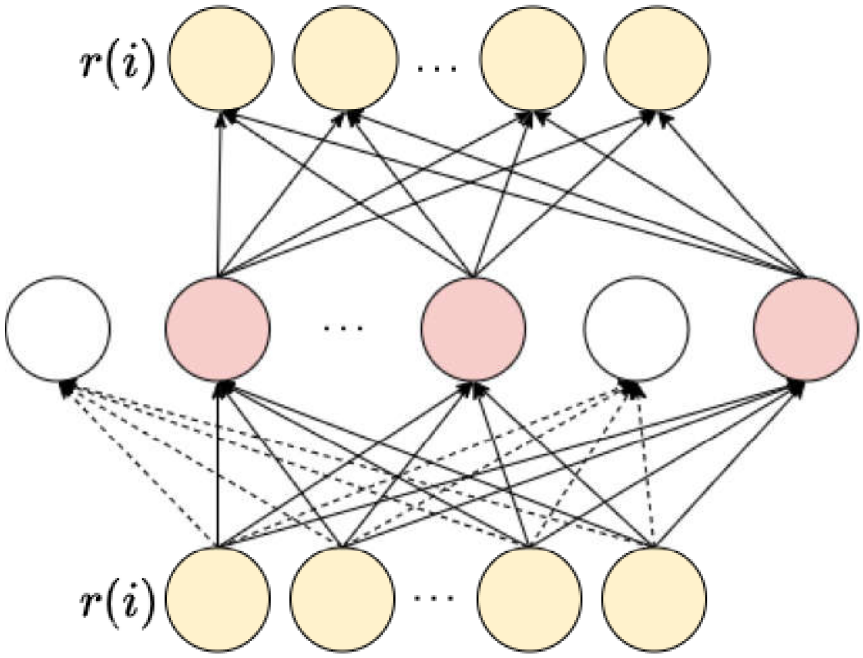
Mặc dù I-Autorec xử lý trực tiếp các dữ liệu đánh giá để đảm bảo tính chính xác trong dự đoán, mô hình này lại bỏ qua những thông tin phụ có giá trị. Để khắc phục hạn chế đó, Autorec đã được thiết kế lại nhằm cho phép tích hợp các vector thông tin liên quan đến nội dung, vốn thường bị bỏ qua trong các mô hình truyền thống. CI-Autorec minh chứng rằng sự khác biệt về kích thước giữa các tầng đầu vào và đầu ra có thể được khai thác để bổ sung thông tin ngoài, từ đó nâng cao hiệu quả so với các phương pháp thông thường (Duong et al., 2023).



Hình 2: Kiến trúc mạng CI-Autorec.

2.3. Mạng tự mã hóa Thưa - Sparse Autoencoder

Sparse Autoencoder là một biến thể đặc biệt của AE, trong đó số lượng nút ẩn được thiết kế lớn hơn số lượng nút đầu vào, đồng thời áp dụng ràng buộc chuẩn L1 nhằm đưa phần lớn các nút ẩn về gần 0 (Ng et al., 2011). Cơ chế này giúp giảm thiểu sự dư thừa trong dữ liệu đầu vào, chỉ giữ lại các đơn vị kích hoạt cần thiết. Nhờ vậy, mô hình trở nên phù hợp để xử lý dữ liệu thưa và dữ liệu có tính dư thừa cao, đồng thời vẫn duy trì được hiệu suất biểu diễn.



Hình 3: Kiến trúc mạng Sparse - Autoencoder

III. Phương pháp nghiên cứu

Phương pháp nghiên cứu được sử dụng trong nghiên cứu này là từ các phân tích và đánh giá những mô hình hiện có, đưa ra các đề xuất cải thiện và tiến hành triển khai các mô hình mới sử dụng dữ liệu thực tế. Mô hình mới được so sánh với những mô hình tham chiếu dựa trên các tiêu chí độ chính xác dự đoán và thời gian thực thi để kiểm nghiệm một cách toàn diện hiệu quả hoạt động. Quá trình này có thể lặp lại nhiều lần để liên tục nâng cao độ chính xác của các mô hình.

IV. Mô hình đề xuất

Trong nghiên cứu trước (Duong et al., 2023), nhóm nghiên cứu đã tiến hành các thử nghiệm thành công với mô hình CI-Autorec - một phương pháp có khả năng giải quyết hiệu quả vấn đề dữ liệu thưa và hiện tượng "khởi đầu lạnh" (cold-start). Bên cạnh đó, các nghiên cứu về Sparse Autoencoder đã cho thấy rằng việc giảm số lượng đơn vị kích hoạt có thể cải thiện hiệu quả cả về mặt tính toán lẫn độ chính xác trong xử lý dữ liệu.

Từ mục tiêu tận dụng những ưu điểm của cả CI-Autorec và Sparse Autoencoder, tác giả đề xuất một mô hình kết hợp hai kỹ thuật này theo hướng cộng hưởng, nhằm nâng cao khả năng diễn giải cũng như độ chính xác trong dự đoán. Tương tự như CI-Autorec - vốn sử dụng thông tin dựa trên nội dung sản phẩm làm đầu vào - mô hình đề xuất bắt đầu bằng việc sử dụng CI-Autorec để xây dựng biểu diễn tiềm ẩn của dữ liệu tại tầng ẩn đầu tiên. Sau đó, một lớp Sparse được thêm vào nhằm tinh chỉnh biểu diễn này. Nhờ sử dụng chuẩn hóa L1, lớp Sparse đảm bảo rằng chỉ một tập nhỏ các đơn vị được kích hoạt trong quá trình tính toán, giúp loại bỏ các tín hiệu nhiễu không cần thiết và nâng cao khả năng diễn giải của mô hình. Cuối cùng, dự đoán đánh giá của người dùng được tính tại tầng đầu ra.

Như minh họa trong Hình 4, mô hình đề xuất bao gồm bốn tầng. Tầng đầu tiên nhận vào vector đặc trưng của mục , trong đó là số lượng đặc trưng. Hai tầng ẩn lần lượt là và , với và là số lượng nơ-ron ẩn tại mỗi tầng. Tầng đầu ra là vector dự đoán đánh giá . Mạng nơ-ron này được mô tả bằng các phương trình sau:

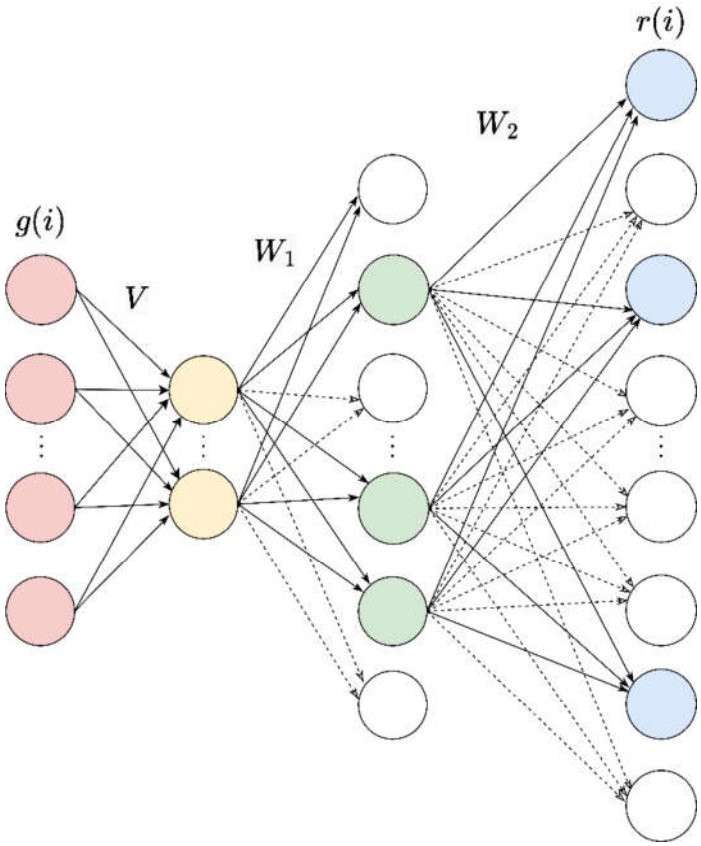
$$h_1 = z_1(W_1 g + b_1) \quad (1)$$

$$h_2 = z_2(W_2 h_1 + b_2) \quad (2)$$

$$\hat{r} = z_3(V h_2 + \mu) \quad (3)$$

Trong đó:

- $W_1 \in R^{k_1 \times f}$, $W_2 \in R^{k_2 \times k_1}$, và $V \in R^{m \times k_2}$ là các ma trận trọng số;
- $b_1 \in R^{k_1}$, $b_2 \in R^{k_2}$ và $\mu \in R^m$ là các vector độ lệch (bias);
- $z_1(\cdot)$, $z_2(\cdot)$, và $z_3(\cdot)$ là các hàm kích hoạt phi tuyến.



Hình 4: Kiến trúc mô hình đề xuất.

Hàm mất mát của mô hình đề xuất không được tính theo sai số tái tạo giữa đầu vào và đầu ra như các mô hình AE truyền thống, mà được xây dựng dựa trên sai số bình phương giữa giá trị đánh giá dự đoán và giá trị thực tế đã quan sát:

$$\mathcal{L}(r, \hat{r}) = \|r - \hat{r}\|_O^2 = \|r - z_3(Vh_2 + \mu)\|_O^2 \quad (4)$$

Trong đó biểu thị rằng chỉ các đánh giá đã được quan sát mới được sử dụng trong quá trình tính toán hàm mất mát.

V. Kết quả thực nghiệm

5.1. Tập dữ liệu

*Đ*ể đánh giá hiệu quả của mô hình đề xuất, hai tập dữ liệu chuẩn là **MovieLens 20M** (Harper and Konstan, 2016) *và* **BookGenome** (Kotkov et al., 2022) được sử dụng.

• **MovieLens 20M**, được phát hành bởi GroupLens vào năm 2015, bao gồm 20.000.263 đánh giá. Phiên bản cập nhật năm 2016 bổ sung thêm 465.564 thẻ (tag) được gán cho 27.278 bộ phim bởi 138.493 người dùng (tất cả đều đã đánh giá ít nhất 20 phim).

• **BookGenome**, được phát hành vào năm 2021, chứa 5.152.656 lượt đánh giá trên 9.374 đầu sách, đến từ 350.332 người dùng.

Dữ liệu Tag Genome, do GroupLens cung cấp, đo lường mức độ một bộ phim phản ánh một đặc trưng nào đó (thông qua các thẻ) theo thang điểm từ 0 đến 1. Các điểm số này được tính toán dựa trên dữ liệu người dùng như thẻ gán, đánh giá và nhận xét văn bản (Harper and Konstan, 2016).

Tập dữ liệu MovieLens với Tag Genome cung cấp siêu dữ liệu phong phú cho từng bộ phim, bao gồm các thẻ mô tả nhiều khía cạnh như thể loại, nội dung, bối cảnh, cảm xúc, phong cách và chủ đề. Ví dụ, một bộ phim có thể được gán các thẻ như "hành động", "hài lãng mạn", "khoa học viễn tưởng", "kết thúc bất ngờ", hay "trưởng thành".

Tương tự, BookGenome mô tả sách thông qua hệ thống thẻ phong phú, thể hiện nội dung và phong cách của từng tác phẩm, bao gồm: thể loại, chủ đề, lối hành văn, cấu trúc, loại nhân vật,... Ví dụ, một quyển sách có thể mang các thẻ như "trinh thám", "tiểu thuyết lịch sử", "triết lý", "phiêu lưu sử thi", hay "xoay quanh nhân vật".

Vì mô hình đề xuất phụ thuộc vào thông tin từ Tag Genome, dữ liệu gốc được xử lý sơ bộ bằng cách loại bỏ các phim không có thông tin thẻ. Đồng thời, chỉ giữ lại những phim và người dùng có ít nhất 20 đánh giá. Bảng 1 *tóm tắt kết quả xử lý:*

Bảng 1: Tổng quan tập dữ liệu trước và sau xử lý

Tập dữ liệu	# Đánh giá	# Người dùng	# Mục (item)	Mức độ thưa
MovieLens 20M				
Tập dữ liệu gốc	20.000.263	138.493	27.278	99,47%
Tập dữ liệu sau xử lý	19.793.342	138.185	10.239	98,97%
BookGenome				
Tập dữ liệu gốc	5.152.656	350.332	9.374	99,99%
Tập dữ liệu sau xử lý	3.799.864	60.717	9.374	99,98%

5.2. Phương pháp đánh giá

Hiệu suất của mô hình được đánh giá dựa trên các chỉ số sau:

Sai số căn trung bình bình phương (Root Mean Squared Error - RMSE): đo lường sai số giữa giá trị đánh giá dự đoán và thực tế, càng nhỏ càng tốt

$$\text{RMSE} = \sqrt{\sum_{u,i \in \text{TESTSET}} (\widehat{r_{ui}} - r_{ui})^2 / |\text{TESTSET}|} \quad (5)$$

Trong đó, $\widehat{r_{ui}}$ là đánh giá dự đoán của người dùng u cho mục i; r_{ui} là đánh giá thực tế và |TESTSET| là số lượng mẫu trong tập kiểm tra.

• ***Độ chính xác tại vị trí K (P@K)*** *và* ***Độ bao phủ tại vị trí K (R@K):*** P@K là tỉ lệ mục liên quan trong top-K mục được đề xuất, còn R@K là tỉ lệ mục liên quan được đề xuất trong top-K trên tổng số mục liên quan.

Dựa trên ma trận phân loại trong Bảng 2, hai chỉ số này được tính như sau:

$$\text{Precision} = \frac{\#\text{tp}}{\#\text{tp} + \#\text{fp}} \quad (6)$$

$$\text{Recall} = \frac{\#tp}{\#tp + \#fn} \quad (7)$$

Bảng 2: Ma trận phân loại trong hệ thống gợi ý

	Liên quan	**Không liên quan**
Đề xuất	TP (true positive)	FP (false positive)
Không đề xuất	FN (false negative)	TN (true negative)

5.3. Mô hình so sánh và thiết lập thực nghiệm

Mô hình đề xuất được so sánh với các mô hình cơ sở phổ biến sau:

• **kNN-Content** (Duong et al., 2019): mô hình lai dựa trên hàng xóm, sử dụng PCC-genome làm thước đo tương đồng, áp dụng các vector nội dung.

• **SVD**: mô hình phân rã ma trận với 40 nhân tố ẩn, học với tốc độ 0.002 trong 100 vòng lặp.

• **FM**$_{\text{genome}}$ (Rendle, 2010): vector đặc trưng gồm mã hóa one-hot của người dùng và phim, thể loại phim, và điểm genome. Mô hình học với bậc đa thức trong 50 vòng.

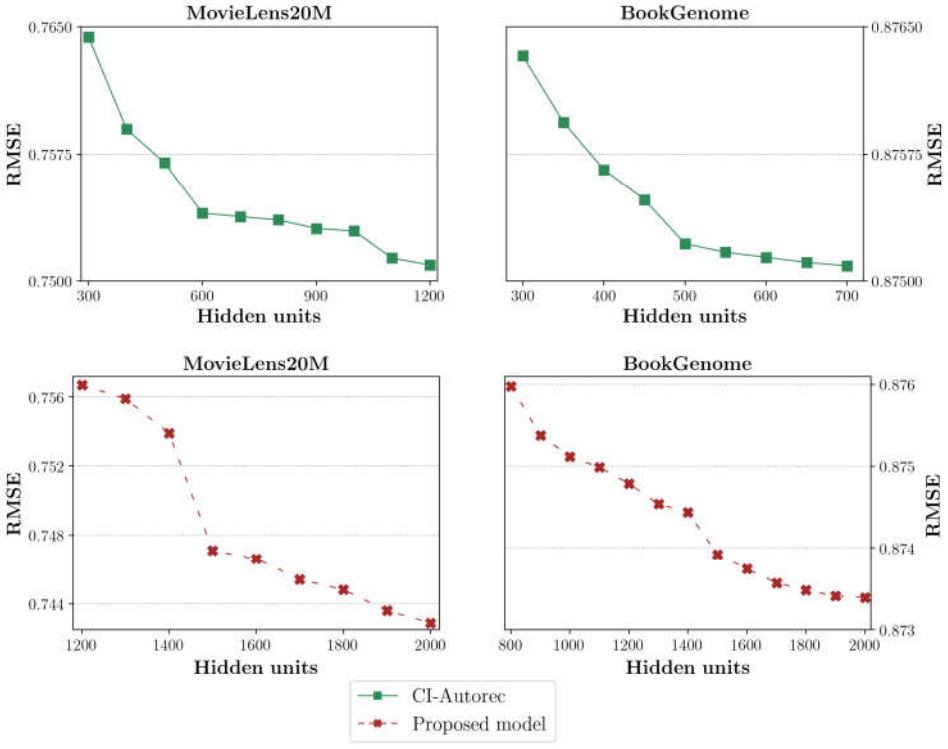
• **I-Autorec** (Sedhain et al., 2015): mô hình 3 lớp với 600 nơ-ron ẩn, dùng cặp hàm kích hoạt (Identity, Sigmoid).

• **CI-Autorec** (Duong et al., 2023): sử dụng điểm genome làm đầu vào, các siêu tham số giữ nguyên như I-Autorec.

Sau xử lý, mỗi tập dữ liệu được chia thành 80% huấn luyện và 20% kiểm tra. Trong tập huấn luyện, 10% được tách để xác thực. Các siêu tham số được tinh chỉnh riêng cho từng mô hình nhằm đảm bảo so sánh công bằng.

5.4. Kết quả thực nghiệm

Để kiểm tra tác động của số lượng các vector ẩn lên hiệu suất của Mô hình đề xuất, tác giả đã thực hiện các thí nghiệm với các số lượng đơn vị ẩn khác nhau ở các tầng nén. Ban đầu, giá trị được kiểm tra trong phạm vi (với tập dữ liệu MovieLens20M) và (với tập dữ liệu BookGenome) với CI-Autorec nhằm xác định độ dài vector tối ưu cho tầng nén đầu tiên. Sau khi cố định , tác giả tiếp tục đánh giá trong phạm vi từ (đối với MovieLens20M) và từ (đối với BookGenome).



Hình 5: Kết quả thực nghiệm với số lượng đơn vị ẩn khác nhau ở các tầng ẩn.

Kết quả thực nghiệm được trình bày trong Hình 5 cho thấy tỷ lệ lỗi tỷ lệ thuận với số lượng đơn vị ẩn. Với mỗi tập dữ liệu, mô hình đạt đến điểm bão hòa RMSE tại các giá trị ***và*** khác nhau. Các kết quả thực nghiệm trong Chương này được thực hiện với tập được trình bày trong Bảng 3.

Bảng 3: Cấu hình tham số ẩn sử dụng trong thí nghiệm

Tập dữ liệu		
MovieLens	600	1500
BookGenome	500	1500

Thực nghiệm trong nghiên cứu này tập trung vào việc đánh giá và so sánh hiệu suất của nhiều mô hình cơ sở khác nhau. Các mô hình này được xây dựng bằng cách sử dụng nhiều thư viện và khung làm việc (framework) đa dạng, phổ biến trong lĩnh vực học máy và học sâu. Mục tiêu chính của quá trình thực nghiệm là xác định mô hình đạt hiệu quả cao nhất thông qua việc giảm thiểu tỷ lệ lỗi hoặc tối ưu hóa các chỉ số đánh giá như Precision và Recall. Đáng chú ý, tác giả chủ động lựa chọn tập trung vào các chỉ số về độ chính xác và tỷ lệ lỗi để phản ánh khả năng dự đoán của các mô hình, trong khi đó yếu tố thời gian cần thiết cho quá trình huấn luyện và dự đoán được loại khỏi phạm vi phân tích. Do đó, Bảng 4 chỉ trình bày các chỉ số liên quan trực tiếp đến hiệu quả dự đoán, nhằm đảm bảo tính khách quan, nhất quán và công bằng trong quá trình đánh giá mô hình.

Bảng 4: So sánh hiệu năng giữa mô hình đề xuất và các mô hình cơ sở

	MovieLens20M			***BookGenome***		
Mô hình	**RMSE**	**P@5**	**R@5**	**RMSE**	**P@5**	**R@5**
kNN-Content	0.7885	0.8008	0.4362	0.8845	0.7662	0.7386
SVD	0.7922	0.8005	0.4322	0.8942	0.7644	0.7371
FM_{genome}	0.7918	0.7993	0.4316	0.8956	0.7617	0.7332
I-Autorec	0.7761	0.8016	0.4341	0.8766	0.7655	0.7392
CI-Autorec	0.7540	0.8115	0.4381	0.8752	0.7743	0.7457
Mô hình đề xuất	**0.7471**	**0.8155**	**0.4388**	**0.8739**	**0.7774**	**0.7482**

Các kết quả thực nghiệm cho thấy khả năng dự đoán của Mô hình đề xuất hiệu quả hơn so với các mô hình cơ sở. Đáng chú ý, hiệu suất của mô hình được tối ưu khi RMSE giảm thiểu và Precision và Recall được tối đa hóa. Mô hình đề xuất cho thấy sự cải thiện đáng kể, dao động từ khoảng đến cải thiện RMSE và tăng chỉ số Precision/Recall từ ***đến*** so với các mô hình cơ sở:

- RMSE giảm từ 5,25% đến 1,20%, Precision/Recall cải thiện khoảng 0,59% - 1,83% so với mô hình kNNContent.
- RMSE giảm từ 5,69% đến 2,27%, Precision/Recall cải thiện khoảng 1,51% - 1,87% so với mô hình SVD.
- RMSE giảm từ 5,64% đến 2,42%, Precision/Recall cải thiện khoảng 1,66% - 2,06% so với mô hình FM_{genome}.
- RMSE giảm từ 3,74% đến 0,31%, Precision/Recall cải thiện khoảng 1,08% - 1,73% so với mô hình I-Autorec.
- RMSE giảm từ 0,92% đến 0,15%, Precision/Recall cải thiện khoảng 0,15% - 0,49% so với mô hình CI-Autorec.

Những kết quả đạt được từ quá trình thực nghiệm đã chứng minh hiệu quả rõ rệt của mô hình đề xuất trong việc cải thiện độ chính xác dự đoán cho các Hệ thống gợi ý, đồng thời góp phần thúc đẩy sự phát

triển của các phương pháp học máy hiện đại. Cụ thể, việc tích hợp thêm một lớp Sparse vào kiến trúc mô hình đã mang lại những cải thiện đáng kể, thể hiện qua các chỉ số đánh giá như tỉ lệ lỗi RMSE và hiệu suất trong các tác vụ xếp hạng top-k sản phẩm. Những kết quả này làm nổi bật vai trò của lớp Sparse trong việc tăng cường khả năng học biểu diễn và khai thác thông tin tiềm ẩn trong dữ liệu. Nhờ đó, mô hình có thể phát hiện và tận dụng hiệu quả các mối quan hệ ẩn giữa các thẻ (tags), từ đó xây dựng được biểu diễn đặc trưng tốt hơn cho từng mục tiêu dự đoán.

VI. Kết luận

Trong nghiên cứu này, tác giả đã đề xuất một mô hình lai ghép giữa CI-Autorec và Sparse Autoencoder nhằm nâng cao độ chính xác của các hệ thống gợi ý dựa trên Autorec. Điểm mới của mô hình là bổ sung một lớp Sparse ngay trước tầng đầu ra, giúp khai thác hiệu quả hơn mối quan hệ tiềm ẩn giữa các đặc trưng đồng thời giảm thiểu nhiễu. Kết quả thực nghiệm trên hai tập dữ liệu chuẩn MovieLens 20M và BookGenome cho thấy mô hình không chỉ vượt trội hơn so với các phương pháp cơ sở như SVD, FMgenome, I-Autorec và CI-Autorec, mà còn đạt mức cải thiện đến 5,96% về RMSE và 2,06% ở các chỉ số top-k (Precision/Recall). Điều này chứng minh khả năng học biểu diễn mạnh mẽ và độ ổn định cao của mô hình khi áp dụng trên dữ liệu lớn và thưa. Hướng nghiên cứu trong tương lai sẽ tập trung vào các khía cạnh sau: (1) Mở rộng mô hình để hỗ trợ thêm các loại dữ liệu không có cấu trúc như văn bản hoặc hình ảnh; (2) Tối ưu hóa mô hình về mặt hiệu năng tính toán nhằm phù hợp hơn với các ứng dụng thời gian thực; (3) Khảo sát khả năng áp dụng mô hình cho các hệ thống gợi ý theo ngữ cảnh (context-aware recommendation) và gợi ý tuần tự (sequential recommendation).

Tài liệu tham khảo:

[1]. A. Ng et al., “Sparse autoencoder,” CS294A Lecture notes, vol. 72, no. 2011, pp. 1-19, 2011.

[2]. B. D. Son, N. T. Hoa, T. V. Chien, W. Khalid, M. A. Ferrag, W. Choi, and M. Debbah, “Adversarial attacks and defenses in 6g network-assisted iot systems,” *IEEE Internet of Things Journal,* 2024, vol. 11, no. 11, pp. 19 168-19 187.

[3]. D. Kotkov, A. Medlar, A. Maslov, U. R. Satyal, M. Neovius, and D. Glowacka, “The tag genome dataset for books,” in *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval, 2022,* pp. 353-357.

[4]. F. M. Harper and J. A. Konstan, “The movielens datasets: History and context,” *Acm transactions on interactive intelligent systems (tiis)*, 2016, vol. 5, no. 4, p. 19.

[5]. G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *science*, 2006, vol. 313, no. 5786, pp. 504-507.

[6]. G. Guo, J. Zhang, and D. Thalmann, “Merging trust in collaborative filtering to alleviate data sparsity and cold start,” *Knowledge-Based Systems*, 2014,vol. 57, pp. 57-68.

[7]. J. Niu, L. Wang, X. Liu, and S. Yu, “Fuir: Fusing user and item information to deal with data sparsity by using side information in recommendation systems,” *Journal of Network and Computer Applications,* 2016, vol. 70, pp. 41-50.

[8]. K. Hornik, M. Stinchcombe, and H. White, “Multilayer feedforward networks are universal approximators,” *Neural networks,* 1989, vol. 2, no. 5, pp. 359-366.

[9]. N. Duong Tan, T. A. Vuong, D. M. Nguyen, and Q. H. Dang, "Utilizing an autoencoder-generated item representation in hybrid recommendation system," *IEEE Access*, 2020, vol. PP, pp. 1-1.

[10]. N. Duong Tan, T. Pham, T. Do, T. Dinh, and M. Tran, "A generalized autorec framework applying content-based information for resolving data sparsity problem," 12 2023, pp. 181-188.

[11]. N. Idrissi and A. Zellou, "A systematic literature review of sparsity issues in recommender systems," *Social Network Analysis and Mining,* 2020, vol. 10, no. 1, pp. 1-23.

[12]. N. T. Hoa, D. Van Dai, L. H. Lan, N. C. Luong, D. Van Le, and D. Niyato, "Deep reinforcement learning for multi-hop offloading in uavassisted edge computing," *IEEE Transactions on Vehicular Technology,* 2023, vol. 72, no. 12, pp. 16 917-16 922.

[13]. M. Ranjbar, P. Moradi, M. Azami, and M. Jalili, "An imputation-based matrix factorization method for improving accuracy of collaborative filtering systems," *Engineering Applications of Artificial Intelligence,* 2015, vol. 46, pp. 58-66.

[14]. S. Rendle, "Factorization machines," in *2010 IEEE International Conference on Data Mining. IEEE, 2010,* pp. 995-1000.

[15]. S. Sedhain, A. K. Menon, S. Sanner, and L. Xie, "Autorec: Autoencoders meet collaborative filtering," in *Proceedings of the 24th International Conference on World Wide Web. ACM, 2015*, pp. 111-112.

[16]. S. Zhang, L. Yao, X. Xu, S. Wang, and L. Zhu, "Hybrid collaborative recommendation via semi-autoencoder," 2017.

[17]. T. N. Duong, T. N. Cao, T. G. Do, T. D. Mai, and M. H. Tran, "A scalable recommendation system with hybrid similarity matrix using accelerated particle swarm optimization," in *2023 International Conference on Advanced Technologies for Communications (ATC), 2023*, pp. 480-487.

[18]. T. N. Duong, T. G. Do, T. N. Cao, and M. H. Tran, "User-item correlation in hybrid neighborhood-based recommendation system with synthetic user data," in *2022 IEEE Ninth International Conference on Communications and Electronics (ICCE). IEEE, 2022,* pp. 176-181.

[19]. T. N. Duong, V. D. Than, T. H. Tran, T. H. A. Pham, H. N. Tran et al., "A practical solution to the acm recsys challenge 2018," in *2018 5th NAFOSTED Conference on Information and Computer Science (NICS). IEEE, 2018,* pp. 341-343.

[20]. T. N. Duong, V. D. Than, T. A. Vuong, T. H. Tran, Q. H. Dang, D. M. Nguyen, and H. M. Pham, "A novel hybrid recommendation system integrating content-based and rating information," in *International Conference on Network-Based Information Systems. Springer, 2019*, pp. 325-337.

[21]. X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in *Proceedings of the 26th international conference on world wide web. International World Wide Web Conferences Steering Committee, 2017*, pp. 173-182.

[22]. X. He and T.-S. Chua, "Neural factorization machines for sparse predictive analytics," in *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval, 2017,* pp. 355-364.

[23]. X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Advances in artificial intelligence*, 2009, vol. 2009,.

[24]. Y. Ouyang, W. Liu, W. Rong, and Z. Xiong, "Autoencoder-based collaborative filtering," in *International Conference on Neural Information Processing.* Springer, 2014, pp. 284-291.

DESIGN OF A GENERALIZED AUTOREC FRAMEWORK FOR RECOMMENDER SYSTEMS

Pham Thi Thu Trang[2]

Abstract: *In recent years, neural network-based collaborative filtering has achieved notable progress in delivering personalized recommendations. CI-AutoRec enables the incorporation of content-based information to enhance recommendation quality, while Sparse Autoencoders have proven effective in generalization and reducing computational costs without compromising accuracy. In this paper, we introduce a novel hybrid framework that integrates the content-aware capability of CI-AutoRec with the representational power of Sparse Autoencoders for rating prediction. Experimental results demonstrate that our approach consistently outperforms state-of-the-art hybrid models in both accuracy and efficiency, highlighting its potential as a robust solution for next-generation recommender systems.Bottom of Form*

Keywords: *recommender systems, deep learning, autoencoders, content-based information, sparse data*

[2] Hanoi Open University