PROJECT 1B: DỰ ĐOÁN MỐI QUAN HỆ HỢP TÁC GIỮA CÁC DIỄN VIÊN ĐIỆN ẢNH VIỆT NAM

1st Bùi Minh Huy, 2nd Trần Lê Vân, 3rd Nguyễn Thị Thanh Tâm and Le Nhat Tung
HUTECH University, Vietnam
huybm.ds@gmail.com, vtran0712004@gmail.com, ttam99852@gmail.com and lenhattung@hutech.edu.vn

Tóm tắt nội dung

Ngành điện ảnh được nhắc tới là một hệ sinh thái phức tạp, trong đó mối liên hệ của các diễn viên giữ vai trò trung tâm. Trong khi nhiều nghiên cứu quốc tế đã phân tích mạng hợp tác diễn viên ở các nền điện ảnh lớn như Hollywood hay Bollywood nhưng các nghiên cứu tương tự trong bối cảnh Việt Nam vẫn còn hạn chế. Để lấp đầy khoảng trống này, Nghiên cứu này tập trung dự đoán mối quan hệ hợp tác giữa các diễn viên điện ảnh Việt Nam dựa trên dữ liệu phim chiếu rạp từ năm 2020 đến nay. Mỗi diễn viên được mô hình hóa bằng một nút và các mối quan hệ hợp tác được biểu diễn bằng một cạnh trong mạng xã hội điện ảnh. Dựa trên cơ sở mạng này, các phương pháp dự đoán liên kết được áp dụng, bao gồm các chỉ số topo (Common Neighbors, Adamic–Adar, Resource Allocation), các mô hình học máy (Logistic Regression, Random Forest, GBDT) và học sâu (Node2Vec, GCN). Kết quả cho thấy Logistic Regression đạt kết quả tốt nhất trong LCC Split (AUC-ROC 0.996, AUC-PR 0.995, AP 0.995) trong khi Random Forest nổi bật nhất ở Shuffle Split với AUC-ROC 0.994 và AUC-PR 0.996. Những phát hiện này không chỉ mở rộng nghiên cứu học thuật về mạng xã hội mà còn cung cấp hàm ý thực tiễn về phát triển điện ảnh Việt Nam.

Từ khóa

Social Network Analysis, Community Detection, Actor Collaboration, Vietnamese Cinema, Complex NNetworks, Link Prediction

I. Giới thiệu

Trong bối cảnh hội nhập và phát triển, ngành công nghiệp điện ảnh Việt Nam đang từng bước khẳng định vị thế, không chỉ là một phương tiện giải trí mà còn là một lĩnh vực kinh tế - văn hóa năng động. Các tác phẩm điện ảnh đã và đang đóng một vai trò quan trọng trong việc hình thành và ảnh hưởng đến các khía cạch của cuộc sống, không chỉ cuốn hút người xem bởi nội dung đa dạng và khả năng kể chuyện sinh động mà còn mang đến những trải nghiệm thẩm mỹ phong phú, những phút giây giải trí và còn là công cụ phản ánh, lưu giữ, lan tỏa bản sắc văn hóa của quốc gia. Trong 5 năm gần đây, điện ảnh đã có sự phát triển mạnh mẹ cả về số lượng lẫn chất lượng của các tác phẩm, qua đó càng khẳng định vị trí quan trọng trong đồi sống tinh thần của khán giả cũng như như trong chiến lược phát triển công nghệp văn hóa đất nước. Trong quá trình sản xuất, mỗi một bộ phim điện ảnh được tạo nên từ sự kết hợp của nhiều cá nhân và tập thể với nhiều vai trò khác nhau. Trong đó, các diễn viên tham gia diễn xuất đóng vai trò trung tâm, bởi chính họ là người trực tiếp thể hiện câu chuyện và kết nối tác phẩm tới khán giả. Trong hệ sinh thái điện ảnh, mối quan hệ hợp tác giữa các diễn viên tưởng chừng là riêng lẻ nhưng thức chất khi kết hợp lại sẽ tạo nên một mạng xã hội phức tạp với cấu trúc và động lực riêng [1]. Mạng lưới này hoạt động liên tục, trong các lựa chọn hợp tác không chỉ phản ánh sự hợp tác về mặt nghệ thuật hay chiến lược của nhà sản xuất, mà còn dần định hình nên các dòng chảy sáng tạo và xu hướng của toàn ngành. Chính vì vậy, việc nghiên cứu các mối quan hệ hợp tác giữa các diễn viên có thể mang lại những hiểu biết về cách ngành công nghiệp điện ảnh hoạt động cũng như phản ánh sự gắn kết nghề nghiệp, những xu hướng cộng đồng và phân hóa trong hệ sinh thái điện ảnh.

Để giải mã các cấu trúc mạng hợp tác của các diễn viên thì phương pháp phân tích mạng xã hội đã được đề xuất khi cho phép khám phá cấu trúc và động lực hình thành các mối quan hệ trong nhiều hệ thống phức tạp. Thì một câu hỏi tiếp theo được đặt ra mang tính chất thực tiễn hơn là liệu trong tương lai những cặp diễn viên nào có khả năng sẽ kết hợp với nhau? Việc dự đoán mối quan hệ hợp tác trong tương lai không chỉ mở rộng phạm vi của phân tích mạng xã hội mà còn đem lại những ứng dụng quan trọng cho ngành công nghiệp điện ảnh. Ta gọi đây là bài toán dự đoán liên kết (link prediction) trong mạng xã hội. Một hướng nghiên cứu quan trọng trong phân tích mạng xã hội, nhằm dự đoán sự hình thành của các cạnh mới dựa trên cấu trúc và đặc trưng của mạng hiện tại [2]. Các nghiên cứu quốc tế cho thấy link prediction có tính ứng dụng rộng rãi như từ gợi ý kết bạn trên mạng xã hội đến mô hình hóa hợp tác trong ngành điện ảnh . Đặc biệt, các nghiên cứu gần đây về mạng hợp tác diễn viên cho thấy các chỉ số có thể phản ánh xác suất hình thành các mối quan hệ hợp tác mới của các diễn viên [3]. Tuy nhiên các nghiên cứu này chưa được áp dụng mạnh mẽ đối với bối cảnh điện ảnh Việt Nam, đã tạo ra một khoảng trống nghiên cứu đáng kể.

Xuất phát từ thực tiễn trên, nghiên cứu này được thực hiện với mục tiêu xây dựng mạng hợp tác diễn viên điện ảnh Việt Nam dựa trên các bộ phim chiếu rạp từ năm 2020 đến nay, đồng thời áp dụng các phương pháp dự đoán liên kết nhằm xác định các cặp diễn viên có tiềm năng hợp tác trong tương lai. Phương pháp nghiên cứu kết hợp giữa các chỉ số tương đồng

topo (node similarity measures), các mô hình xác suất (probabilistic models), các mô hình học máy có giám sát (supervised learning). Bên cạnh đó, các phương pháp học sâu hiện đại như Node2Vec và Graph Convolutional Network (GCN) cũng được thử nghiệm để đánh giá hiệu năng và khả năng khái quát hóa trong bối cảnh dữ liệu điện ảnh Việt Nam. Nghiên cứu hướng tới việc không chỉ mở rộng ứng dụng của SNA và link prediction, mà còn đóng góp vào việc hiểu sâu hơn về cấu trúc và động lực hợp tác trong ngành công nghiệp điện ảnh Việt Nam.

A. Động lực và thách thức

Ngành điện ảnh Việt Nam đang phát triển mạnh mẽ với sự gia tăng đáng kể về số lượng và chất lượng các tác phẩm, kéo theo sự mở rộng không ngừng của mạng lưới hợp tác giữa các diễn viên. Tuy nhiên, các nghiên cứu theo hướng mạng lưới hợp tác giữa các diễn viên vẫn còn hạn chế, chủ yếu tập trung vào doanh thu, thể loại hoặc chiến lược truyền thông hoặc dừng lại ở những thống kê mô tả đơn lẻ và phải đối diện với nhiều thách thức. Ta phải đối mặt với việc thu thập dữ liệu vì thông tin về các bộ phim và danh sách diễn viên mặc dù có thể thu thập được nhưng thường phân tán, thiếu chuẩn hóa và xảy ra tình trạng trùng lặp hoặc sai lệch thông tin. Bài toán dự đoán liên kết đòi hỏi phải tạo ra cả tập hợp các cặp diễn viên đã hợp tác (positive samples) và chưa từng hợp tác (negative samples) nhưng trong thực tế, việc sinh negative samples có thể dẫn đến thiên lệch hoặc sai lệch đánh giá nếu không được thiết kế cẩn trọng. Cấu trúc mạng hợp tác của ngành điện ảnh luôn biến động theo thời gian, khi mỗi năm có thêm nhiều bộ phim mới và các mối quan hệ cộng tác thay đổi nhanh chóng. Điều này khiến cho các mô hình dự đoán tĩnh có nguy cơ mất hiệu lực, đòi hỏi phương pháp phải được xem xét trong mối liên hệ với tiến trình phát triển động của mạng. Bản chất phức tạp của sự hợp tác diễn viên không chỉ phản ánh các yếu tố cấu trúc mạng đơn thuần mà còn gắn liền với bối cảnh thực tiễn như thể loại phim, hãng sản xuất hay sự quen biết trong nghề, khiến cho việc diễn giải kết quả cần được đặt trong góc nhìn đa chiều.

Những thách thức nêu trên chính là động lực thúc đẩy nghiên cứu này. Vì vậy, việc thiết kế một phương pháp tiếp cận có hệ thống, bao gồm chuẩn hóa dữ liệu, xây dựng mạng lưới hợp tác đáng tin cậy, áp dụng các thuật toán dự đoán liên kết (từ heuristic đến học máy và học sâu) và diễn giải kết quả trong bối cảnh Việt Nam, được kỳ vọng sẽ góp phần thu hẹp khoảng cách nghiên cứu và mang lại giá trị ứng dụng thực tế cho ngành công nghiệp điện ảnh quốc gia.

B. Đóng góp nghiên cứu

Nghiên cứu này mang lại một số đóng góp quan trọng cho lĩnh vực phân tích mạng xã hội và ứng dụng trong điện ảnh Việt Nam, được tóm tắt như sau:

(i) Xây dựng bộ dữ liệu thực nghiệm toàn diện về mạng hợp tác giữa các diễn viên điện ảnh Việt Nam, bao gồm các bộ phim chiếu rạp từ năm 2020 đến nay. Dữ liệu được thu thập, chuẩn hóa và loại bỏ trùng lặp, đảm bảo tính toàn vẹn và khả năng tái sử dụng trong các nghiên cứu tiếp theo.

(ii) Đề xuất và so sánh có hệ thống nhiều nhóm phương pháp dự đoán liên kết, từ các chỉ số topo truyền thống (Common Neighbors, Jaccard, Adamic–Adar, Resource Allocation, Preferential Attachment) đến các mô hình học máy có giám sát (Logistic Regression, Random Forest, Gradient Boosted Decision Trees) và các mô hình học sâu (Node2Vec + Logistic Regression, Graph Convolutional Network – GCN).

(iii) Thực hiện đánh giá thực nghiệm toàn diện bằng các thước đo AUC-ROC, AUC-PR và Average Precision, kết hợp hai chiến lược chia dữ liệu (LCC Split và Shuffle Split) để kiểm chứng tính ổn định và khả năng tổng quát hóa của mô hình. Kết quả cho thấy các chỉ số Adamic–Adar và Resource Allocation đạt hiệu năng cao nhất trong nhóm heuristic, trong khi Random Forest kết hợp Shuffle Split cho kết quả tối ưu trong nhóm học máy.

C. Tổ chức bài báo

Phần còn lại của bài báo được tổ chức như sau. Mục II trình bày các công trình nghiên cứu liên quan. Mục III mô tả khung phương pháp nghiên cứu. Mục IV cung cấp chi tiết về quá trình xây dựng và xử lý dữ liệu. Mục V mô tả thiết kế thí nghiệm và các tiêu chí đánh giá. Mục VI trình bày kết quả thực nghiệm. Mục VII đưa ra phần kết luận, đồng thời thảo luận những hạn chế còn tồn tại và đề xuất các hướng nghiên cứu trong tương lai.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Phần này cung cấp một đánh giá toàn diện về các phương pháp hiện có, được tổ chức bởi các nguyên tắc cơ bản và phương pháp tính toán của chúng.

A. Khái quát về phân tích mạng xã hội (SNA)

Phân tích mạng xã hội (Social Network Analysis – SNA) từ lâu đã trở thành một trong những phương pháp nền tảng để nghiên cứu, khám phá cấu trúc và động lực hình thành các mối quan hệ trong những hệ thống phức tạp trong nhiều lĩnh vực khác nhau như sinh học, truyền thông xã hội và khoa học. Với khả năng mô hình hóa các các thực thể như nút (nodes) và mối quan hệ giữa chúng là những cạnh (edges), SNA cung cấp cho các nhà nghiên cứu những công cụ mạnh mẽ, cho phép các nhà nghiên cứu nhận diện các đặc trưng quan trọng của mạng lưới chẳng hạn như vai trò trung tâm, mức độ gắn kết, cấu trúc cộng đồng và mô hình phân tầng trong mạng lưới. Các nghiên cứu về mạng phức tạp đã cho thấy rằng nhiều hệ thống xã

hội và tự nhiên đều sở hữu những đặc điểm chung [4] bao gồm tính chất "thế giới nhỏ" (small-world property), sự xuất hiện của các nút trung tâm có ảnh hưởng lớn (hubs) và phân bố bậc theo luật lũy thừa, đã được phát hiện trong nhiều loại mạng. Những phát hiện này khẳng định tính phổ quát của SNA, đồng thời nhấn mạnh tiềm năng ứng dụng của nó trong việc nghiên cứu các lĩnh vực khác nhau và tiềm năng ứng dụng rộng rãi của SNA trong việc phân tích các hệ thống đa dạng.

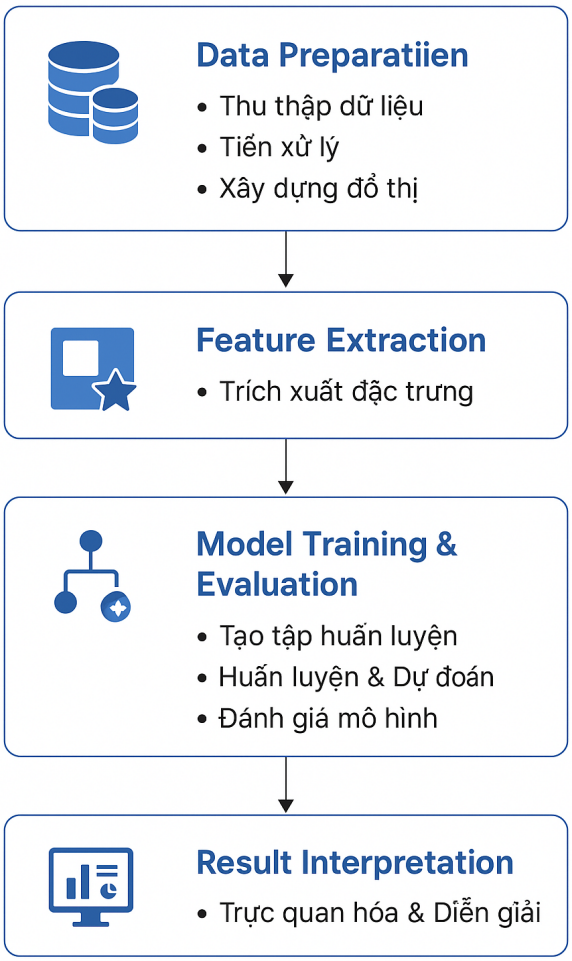
B. Nghiên cứu mạng hợp tác diễn viên trong điện ảnh quốc tế

Trong bối cảnh ngành điện ảnh, mạng hợp tác giữa các diễn viên đã trở thành một vấn đề nghiên cứu điển hình và thu hút nhiều sự quan tâm của cộng đồng khoa học quốc tế. Tiêu biểu là Park và Newman (2005), đã chứng minh rằng mạng hợp tác diễn viên không chỉ phản ánh cơ chế sản xuất phim mà còn thể hiện các đặc tính phổ quát của mạng phức tạp, như cấu trúc "thế giới nhỏ" và sự tồn tại của các nút trung tâm có ảnh hưởng lớn. Tiếp nối hướng nghiên cứu này, các công trình gần đây tại Bollywood đã mở rộng sang việc không chỉ mô tả cấu trúc mà còn dự đoán khả năng hình thành các mối quan hệ hợp tác mới trong tương lai. Bhat và Rohe (2021) chứng minh rằng các chỉ số dựa trên cấu trúc topo, chẳng hạn như Common Neighbors, Jaccard coefficient hay Adamic-Adar, có khả năng phản ánh xác suất hợp tác của các diễn viên trong các dự án sắp tới [2]. Các nghiên cứu mới hơn cũng chứng minh rằng việc kết hợp các đặc trưng topo với mô hình học máy giúp nâng cao khả năng dự đoán trong mạng hợp tác điện ảnh [3]. Đáng chú ý ở đây là các phân tích không chỉ dừng lại ở việc mô tả cấu trúc, mà còn đi sâu vào việc lý giải các yếu tố văn hóa-xã hội, các yếu tố chi phối sự hình thành mạng lưới, tạo ra các cụm hợp tác bền chặt cũng như là dự đoán các liên kết. Những kết quả đã khẳng định rằng mạng lưới hợp tác của các diễn viên trong giới nghệ thuật không chỉ là một cấu trúc kỹ thuật mà còn là sản phẩm của bối cảnh lịch sử, văn hóa và cơ chế vận hành đặc thù của mỗi nền điện ảnh.

C. Phương pháp dự đoán liên kết(Link Prediction)

Các nghiên cứu quốc tế đã chứng minh được giá trị của phân tích mạng xã hội đối với sự hợp tác trong điện ảnh, sự thiếu vắng các công trình tương tự tại Việt Nam tạo ra một khoảng trống cả về học thuật lẫn thực tiễn. Bài toán dự đoán liên kết là một hướng nghiên cứu cốt lõi trong phân tích mạng xã hội, nhằm dự đoán các khả năng xuất hiện của các cạnh mới giữa các nút dựa trên cấu trúc hiện có sẵn [5]. Nghiên cứu này nhằm lấp đầy những khoảng trống đó bằng cách xây dựng và phân tích mạng hợp tác của diễn viên điện ảnh Việt Nam dựa trên dữ liệu phim chiếu rạp từ năm 2020 cho đến nay. Các phương pháp dự đoán lên kết được phân thành ba nhóm chính: Nhóm heuristic sử dụng các thước đo cục bộ giữa hai nút như Common Neighbors, Jaccard, Adamic-Adar và Resource Allocation [2], nhóm học máy có giám sát biểu diễn các cặp nút bằng vector đặc trưng và tiến hành huấn luyện các mô hình phân loại như Logistic Regression, Random Forest [6], nhóm mô hình học sâu sử dụng các mô hình như Node2Vec, GraphSAGE hoặc Graph Convolutional Networks (GCNs) để học biểu diễn nút sau đó tiến hành dự đoán khả năng tồn tại của các liên kết [7]. Các nghiên cứu trên đã so sánh các phương pháp và tìm ra được các mô hình tối ưu.

Công trình không chỉ kế thừa mà còn mở rộng ra các hướng tiếp cận này. Thử nghiệm trên mạng hợp tác diễn viên Việt Nam với hai chiến lược chia dữ liệu (LCC Split và Shuffle Split), đồng thời so sánh hiệu quả giữa các nhóm phương pháp, qua đó xác định tổ hợp mô hình tối ưu cho bài toán dự đoán hợp tác trong điện ảnh Việt Nam. Nghiên cứu không chỉ diễn giải về mặt toán học mà còn diễn giải kết quả trong bối cảnh đặc thù của ngành điện ảnh Việt Nam, nơi các yếu tố văn hóa, lịch sử và tổ chức sản xuất có vai trò quyết định trong việc hình thành mạng lưới. Bằng việc kết hợp phân tích định lượng và diễn giải bối cảnh, nghiên cứu định vị mình như một bước khởi đầu có hệ thống cho việc nghiên cứu cộng đồng trong điện ảnh Việt Nam, đồng thời đóng góp những hàm ý thiết thực cho chiến lược phát triển của ngành công nghiệp văn hóa – nghệ thuật quốc gia.



Hình 1: Quy trình dự đoán liên kết trong mạng diễn viên

III. PHƯƠNG PHÁP NGHIÊN CỨU

Bài nghiên cứu được triển khai theo một khung phương pháp có hệ thống nhằm xây dựng và phân tích mạng hợp tác của các diễn viên điện ảnh Việt Nam.

A. Xây dựng mạng hợp tác diễn viên

Một đồ thị G sẽ được kí hiệu: $G = (V, E)$ với V là tập hợp chứa các đỉnh, và E là tập hợp chứa các cạnh, mỗi cạnh có dạng một cặp giá trị $\{u, v\}$ (có thể được viết thành uv) [8]. Tập hợp đỉnh V của đồ thị G được kí hiệu $V(G)$, tập hợp cạnh được kí hiệu $E(G)$. Dữ liệu được thu thập từ danh sách phim chiếu rạp Việt Nam từ năm 2020 cho đến nay, trong đó mỗi một bộ phim có các thuộc tính như tên phim, đạo diễn, diễn viên, thể loại, ngày khởi chiếu. Tiến hành tiền xử lý để loại bỏ các trùng lặp và chuẩn hóa tên diễn viên thì mỗi diễn viên được mô hình hóa thành một nút $v \in V$, mối quan hệ hợp tác giữa hai diễn viên được biểu diễn bằng một cạnh $e = (u, v) \in E$.

Để phản ánh mức độ hợp tác của mỗi diễn viên, mỗi cạnh được gán một trọng số dựa trên số lần hai diễn viên đó xuất hiện cùng nhau trong các tác phẩm điện ảnh. Cách tiếp cận này đảm bảo rằng mạng hợp tác phản ánh trung thực cường độ hợp tác thực tế giữa các diễn viên. Trong nghiên cứu này, chúng tôi sử dụng hai chiến lược chia tách dữ liệu phổ biến là chỉ giữ lại thành phần liên thông lớn nhất (Largest Connected Component – LCC) để đảm bảo tính toàn vẹn cấu trúc và tránh nhiễu trong quá trình học mô hình. Shuffle Split chia ngẫu nhiên các cạnh thành tập huấn luyện và kiểm tra theo tỷ lệ 80/20 để đánh giá khả năng tổng quát của mô hình. Mặc dù trong thực tế có thể định nghĩa trọng số cạnh như cường độ hợp tác (số lần hai diễn viên cùng xuất hiện trong các phim khác nhau), tuy nhiên trong tập dữ liệu này tỷ lệ cạnh có trọng số lớn hơn 1 chỉ chiếm 2.76% trên tổng số $|E| = 3405$. Do tỷ lệ này nhỏ và có thể gây nhiễu trong quá trình huấn luyện, toàn bộ các thí nghiệm trong nghiên cứu được tiến hành trên đồ thị không trọng số, trong đó mỗi cạnh chỉ phản ánh sự tồn tại của mối quan hệ hợp tác (0 hoặc 1). Để đảm bảo tính toàn vẹn của cấu trúc mạng, nghiên cứu chỉ giữ lại thành phần liên thông lớn nhất

Largest Connected Component – LCC) của đồ thị cho quá trình huấn luyện và kiểm định. Ngoài ra, chiến lược Shuffle Split được áp dụng để chia ngẫu nhiên các cạnh thành tập huấn luyện và kiểm tra theo tỷ lệ 80/20, nhằm đánh giá khả năng tổng quát hóa của mô hình.

B. Đặc trưng cấu trúc mạng

Sau khi mạng hợp tác $G = (V, E)$ được xây dựng hoàn chỉnh, chúng tôi tiến hành phân tích và trích xuất các đặc trưng cấu trúc để mô tả tính chất của mạng, phục vụ cho bài toán dự đoán liên kết (Link Prediction). Các đặc trưng này giúp phản ánh mức độ gắn kết, tính tập trung và đặc tính hình học của mạng, đồng thời là cơ sở cho việc tính toán các thước đo tương đồng giữa các cặp diễn viên. Các chỉ số toàn cục (global metrics) được tính gồm:

– Bậc trung bình (Average Degree): Phản ánh mức độ kết nối trung bình của mỗi nút mạng, được xác định bởi công thức:

$$\bar{k} = \frac{2m}{n} \tag{1}$$

trong đó m là số cạnh và n là số nút. Giá trị $\bar{k}$ càng cao cho thấy các diễn viên trong ngành có xu hướng cộng tác rộng rãi, hình thành nhiều mối qian hệ hợp tác đá chiều trong các bộ phim.

– Mật độ mạng (Network Density): đo lường tỷ lệ số lượng kết nối thực tế và số kết nối tối đa có thể có trong mạng, được tính theo:

$$D = \frac{2m}{n(n-1)} \tag{2}$$

Giá trị D nằm trong khoảng $[0, 1]$; mật độ cao biểu thị cho mức độ gắn kết chặt chẽ giữa các diễn viên, mật độ thấp thì mạng hợp tác còn phân tán, với nhiều nhóm nhỏ hoạt động tách biệt.

– Hệ số gom cụm trung bình (Average Clustering Coefficient): Phản ánh khả năng các diễn viên trong cùng một nhóm có xu hướng hợp tác với nhau như thế nào.Được xác định bởi công thức:

$$\bar{C} = \frac{1}{n} \sum_{i=1}^{n} C_i \tag{3}$$

Với C_i là hệ số gom cụm cục bộ của nút i, biểu thị xác suất hai diễn viên cùng hợp tác với một diễn viên thứ ba thì sẽ có quan hệ hợp tác với nhau. Hệ số gom cụm cao cho thấy mạng có cấu trúc "ê-kíp"đặc trưng, nơi các diễn viên thường xuyên hợp tác trong cùng một nhóm.

– Độ dài đường đi trung bình (Average Path Length): hản ánh mức độ gần gũi giữa các diễn viên trong toàn mạng, được tính bằng:

$$L = \frac{1}{n(n-1)} \sum_{i \neq j} d(i, j) \tag{4}$$

Trong đó $d(i, j)$ là độ dài đường đi ngắn nhất giữa nút i và j. Giá trị L thấp biểu thị rằng bất kỳ diễn viên nào cũng có thể kết nối đến người khác chỉ qua một vài bước hợp tác, thể hiện tính chất "thế giới nhỏ"đặc trưng của mạng xã hội.

– Đường kính mạng: được xác định là khoảng cách ngắn nhất dài nhất giữa hai nút bất kỳ, cho biết mức độ lan tỏa cực đại của mạng. Trong mạng diễn viên, đường kính càng nhỏ chứng tỏ khả năng liên kết và trao đổi cơ hội hợp tác trong ngành càng mạnh.

C. Trích xuất đặc trưng liên kết

Trong bài toán dự đoán liên kết, mỗi cặp diễn viên (u, v) được biểu diễn thành một vector đặc trưng cục bộ như Common Neighbors (CN) [9]Theo [3] thì với hai đỉnh x và y, khả năng hình thành một liên kết giữa chúng sẽ xảy ra nếu chũng có một hoặc nhiều bạn chung. Độ đo đơn giản nhất là hệ số trùng lặp hàng xóm được tính trực tiếp bằng các đếm các bạn chung và đánh dấu, Jaccard Coefficient (JC), Adamic/Adar Index (AA), Preferential Attachment (PA), Resource Allocation Index (RA) [10] đây là các chỉ số đo lường mức độ tương đồng về láng giềng giữa hai nút, được chứng minh có hiệu quả cao trong mạng xã hội [5].

– Số lượng láng giềng chung (Common Neighbors - CN)

$$CN(x, y) = |\Gamma(x) \cap \Gamma(y)| \tag{5}$$

Trong đó:

- $\Gamma(x)$: tập láng giềng của node x.
- $\Gamma(y)$: tập láng giềng của node y.

– Hệ số Jaccard (Jaccard Coefficient - JC)

$$JC(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \tag{6}$$

– Chỉ số Adamic-Adar (Adamic-Adar Index - AA)

$$AA(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log(|\Gamma(z)|)} \tag{7}$$

– Chỉ số phân bố tài nguyên (Resource Allocation Index - RA)

$$RA(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{|\Gamma(z)|} \tag{8}$$

– Độ ưu tiên kết nối (Preferential Attachment - PA)

$$PA(x,y) = |\Gamma(x)| \times |\Gamma(y)| \tag{9}$$

– Khoảng cách ngắn nhất (Shortest Path – SP)
Đặc trưng này phản ánh độ gần giữa hai nút trong mạng, được đo bằng độ dài đường đi ngắn nhất giữa hai nút x và y. Công thức được biểu diễn như sau:

$$SP(x,y) = \frac{1}{d(x,y)} \tag{10}$$

Trong đó, $d(x,y)$ là số cạnh trong đường đi ngắn nhất nối giữa hai nút x và y. Giá trị $SP(x,y)$ càng lớn khi hai nút càng gần nhau, thể hiện xác suất hợp tác tiềm năng cao hơn giữa các diễn viên trong mạng. Nếu không tồn tại đường đi giữa x và y, giá trị $SP(x,y)$ được đặt bằng 0.
Các chỉ số này được chứng minh có khả năng mô tả tốt mức độ tương đồng cấu trúc giữa hai nút, qua đó ước lượng xác suất hình thành cạnh mới trong mạng.

D. Chia tách dữ liệu

Để đánh giá hiệu năng của các mô hình một cách toàn diện và đảm bảo tính khách quan, nghiên cứu sử dụng hai chiến lược chia tách dữ liệu là LCC Split và Shuffle Split. Cả hai phương pháp đều tuân theo tỷ lệ chia cơ bản 80/20 giữa tập huấn luyện và tập kiểm tra, song khác nhau ở cách xử lý cấu trúc mạng trong quá trình tách. LCC Split [11] chỉ giữ lại thành phần liên thông lớn nhất (Largest Connected Component – LCC) của mạng để tiến hành chia dữ liệu. Trong quá trình tách, ưu tiên đảm bảo rằng thành phần huấn luyện vẫn duy trì tính liên thông toàn vẹn. Cụ thể, phương pháp cố gắng loại bỏ khoảng 20% số cạnh để tạo tập kiểm tra (test set) trong khi phần 80% còn lại được sử dụng cho huấn luyện. Nếu việc loại bỏ đủ 20% cạnh làm cho mạng huấn luyện bị tách rời, tỷ lệ loại bỏ sẽ được giảm nhẹ để đảm bảo đồ thị huấn luyện vẫn liên thông. Ngược lại, nếu có thể loại bỏ nhiều hơn mà vẫn duy trì tính liên thông, chỉ đúng 20% cạnh được chọn làm tập kiểm tra. Cách tiếp cận này giúp bảo toàn cấu trúc mạng hợp tác của các diễn viên, phản ánh chính xác mối quan hệ hợp tác thực tế trong ngành điện ảnh. Shuffle Split [12] chia ngẫu nhiên các cạnh của mạng thành hai tập: huấn luyện (80%) và kiểm tra (20%). Mỗi cặp diễn viên có quan hệ hợp tác thực tế được gán nhãn positive, trong khi các cặp chưa từng hợp tác được gán nhãn negative theo tỷ lệ cân bằng, nhằm tránh tình trạng mất cân bằng dữ liệu trong quá trình huấn luyện. Phương pháp này cho phép kiểm tra khả năng khái quát hóa của mô hình khi gặp các liên kết hoàn toàn mới, chưa từng xuất hiện trong quá trình học. Việc kết hợp hai chiến lược LCC Split và Shuffle Split giúp nghiên cứu không chỉ so sánh hiệu năng giữa các nhóm mô hình (heuristic, học máy, học sâu) mà còn đánh giá tính ổn định, khả năng tổng quát và độ tin cậy của từng phương pháp trong các cấu trúc mạng khác nhau.

E. Các mô hình sử dụng

Nghiên cứu thực hiện với bốn nhóm mô hình nhằm đánh giá hiệu năng dự đoán liên kết giữa các diễn viên. Mỗi nhóm sẽ có một cách tiếp cận khác nhau từ các chỉ số topo đến các mô hình học máy và học sâu hiện đại. Nhóm đầu tiên là nhóm heuristic sử dụng trực tiếp thông tin cấu trúc cục bộ của mạng để tính toán mức độ tương đồng giữa hai nút. Nhóm phương pháp này có ưu điểm đơn giản, trực quan, không yêu cầu quá trình huấn luyện phức tạp, nhưng hiệu năng có thể bị giới hạn trong các mạng có cấu trúc không đồng nhất. Sử dụng trực tiếp các thước đo chỉ số tương đồng (similarity metrics) được xây dựng nên tư cấu trúc mạng. Các chỉ số bao gồm: Common Neighbors (CN), Jaccard Coefficient (JC), Adamic–Adar Index (AA), Resource Allocation Index (RA), Preferential Attachment (PA) để tính điểm liên kết, được tính từ các đặc trưng cục bộ của mạng và không yêu cầu quá trình huấn luyện mô hình. Mỗi chỉ số phản ánh một giả thuyết khác nhau về cách

các mối quan hệ hợp tác có thể hình thành. Tiếp theo, là nhóm học máy với đặc trưng thủ công (supervised learning on handcrafted features), các cặp diễn viên (u, v) được biểu diễn bởi vector đặc trưng topo (số láng giềng chung, hệ số cụm, bậc nút, khoảng cách ngắn nhất,...) và gán nhãn nhị phân positive/negative. Trên không gian đặc trưng này, ba mô hình được huấn luyện là Logistic Regression (LR) mô hình tuyến tính cơ bản cho phép ước lượng xác suất tồn tại của liên kết, Random Forest (RF) tập hợp nhiều cây quyết định (decision trees) huấn luyện ngẫu nhiên, Gradient Boosted Decision Trees (GBDT) kỹ thuật boosting xây dựng mô hình mạnh hơn bằng cách kết hợp các cây yếu liên tiếp. Nhóm thứ ba là nhóm học biểu diễn (representation learning), học embedding nút từ cấu trúc mạng rồi dùng embedding đó cho dự đoán liên kết. Nghiên cứu sử dụng Node2Vec [13] để sinh vector cho từng diễn viên thông qua các random walk có điều hướng; hai embedding được kết hợp bằng Hadamard product và phân loại bởi Logistic Regression. Cách tiếp cận này tận dụng thông tin bậc cao nhưng vẫn giữ pipeline hai bước học biểu diễn để phân loại. Nhóm cuối cùng là Học sâu end-to-end trên đồ thị (end-to-end deep graph learning), khác với mô hình học sâu Graph Convolutional Network (GCN) [14] học biểu diễn và dự đoán trong cùng một kiến trúc, áp dụng mô hình hai tầng (2-layer GCN) có khả năng tổng hợp thông tin từ các láng giềng bậc cao, trong đó các embedding nút được kết hợp bằng phép nhân Dot Product để dự đoán xác suất liên kết giữa hai diễn viên [6]. Cách tiếp cận end-to-end cho phép tối ưu trực tiếp mục tiêu dự đoán liên kết nhưng nhạy với quy mô và độ thưa của mạng.

F. Đánh giá mô hình

Chúng tôi sử dụng ba thước đo chính để đánh giá kết quả của các mô hình AUC-ROC, AUC-PR và Average Precision (AP). Các thước đo này phản ánh khả năng phân biệt, mức độ ổn định của tổng thể. AUC-ROC (Area Under the Receiver Operating Characteristic Curve) [15] đo độ phân biệt giữa các liên kết thực tế và các liên kết không tồn tại (positive/negative links). Đường cong ROC được biểu diễn bằng mối quan hệ giữa True Positive Rate (TPR) và False Positive Rate (FPR) khi thay đổi ngưỡng phân loại, được xác định bởi công thức:

$$AUC - ROC = \int_0^1 TPR(FPR)\, d(FPR) \tag{11}$$

Trong đó:

$$TPR = \frac{TP}{TP + FN}, \tag{12}$$

$$FPR = \frac{FP}{FP + TN} \tag{13}$$

AUC-ROC càng cao chứng tỏ mô hình càng có khả năng phân tách tốt giữa các cặp diễn viên có hợp tác và không hợp tác.

AUC-PR (Area Under the Precision–Recall Curve) được sử dụng trong bối cảnh dữ liệu mất cân bằng, khi số lượng liên kết thực nhỏ hơn nhiều so với số lượng cặp chưa hợp tác. Đường cong Precision–Recall biểu diễn mối quan hệ giữa Precision và Recall, được mô tả ở (12)–(13). Chỉ số này phản ánh mức độ chính xác của mô hình đối với các liên kết hiếm.

AP (Average Precision) biểu thị giá trị trung bình của Precision tại các mức Recall khác nhau, được tính theo công thức:

$$AP = \sum_n (R_n - R_{n-1}) P_n \tag{14}$$

Trong đó P_n và R_n lần lượt là Precision và Recall tại ngưỡng n. AP cung cấp một đánh giá tổng hợp về hiệu quả xếp hạng (ranking performance) của mô hình.

Từ các thước đo trên ta có thể xác nhận được các phương pháp kết hợp hay phương pháp độc lập là có tối ưu cho bài toán hay không, từ kết quả ta có thể xem được mô hình nào phù hợp nhất cho dự đoán hợp tác giữa các diễn viên trong nền điện ảnh của Việt Nam.

IV. XÂY DỰNG VÀ CHUẨN BỊ DỮ LIỆU

A. Nguồn dữ liệu và cách thu thập

Nghiên cứu sử dụng hai nguồn dữ liệu chính. Thứ nhất là **bộ dữ liệu phim chiếu rạp Việt Nam** giai đoạn 2020–nay, bao gồm 176 tác phẩm với thông tin chi tiết về tên phim, đạo diễn, danh sách diễn viên, thể loại và ngày công chiếu. Thứ hai là **bộ dữ liệu diễn viên Việt Nam** với 618 diễn viên, bổ sung thêm các thông tin về tiểu sử, quốc tịch, quá trình hoạt động nghệ thuật và hình ảnh minh họa. Hai nguồn dữ liệu này là nền tảng để xây dựng mạng cộng tác, trong đó các nút là diễn viên và các cạnh phản ánh quan hệ hợp tác qua từng bộ phim.

Bảng I minh họa cấu trúc dữ liệu phim. Mỗi hàng đại diện cho một bộ phim, trong đó trường "Diễn viên" chứa nhiều tên cần được tách, chuẩn hóa và loại bỏ trùng lặp trong giai đoạn tiền xử lý.

Bảng I: Ví dụ về cấu trúc dữ liệu phim chiếu rạp Việt Nam từ năm 2020

Tên phim	Đạo diễn	Diễn viên	Thể loại	Khởi chiếu
Đôi mắt âm dương	Nhất Trung	Thu Trang, Quốc Trường, NSND Ngọc Giàu, Trung Dân...	Tâm lý, Kinh dị	25/01/2020
30 chưa phải Tết	Nguyễn Quang Huy	NSND Hồng Vân, Trường Giang, Mạc Văn Khoa, Đức Phúc...	Hài	25/01/2020
Gái già lắm chiêu 3	Bảo Nhân, Nam Cito	Ninh Dương Lan Ngọc, NSND Lê Khanh, Jun Vũ...	Hài, Lãng mạn	25/01/2020

B. Thống kê và đặc trưng dữ liệu

Bảng II trình bày các đặc trưng thống kê chính.

Bảng II: Thống kê đặc trưng của mạng cộng tác

Dữ liệu	Số nút	Số cạnh	Bậc trung bình	Hệ số gom cụm	Đường kính	Số cộng đồng
Mạng phim	563	3,405	12.096	0.7907	7	7

Kết quả cho thấy mạng phim có quy mô lớn hơn nhưng mức độ gắn kết cộng đồng thấp hơn, trong khi mạng diễn viên nhỏ hơn nhưng chặt chẽ hơn, với hệ số gom cụm cao.

C. Quy trình tiền xử lý dữ liệu

Để đảm bảo chất lượng, dữ liệu được xử lý theo các bước:

- Làm sạch: Loại bỏ cạnh trùng lặp, cạnh tự nối và các nút cô lập.
- Lọc thành phần: Giữ lại thành phần liên thông lớn nhất để duy trì tính kết nối toàn cục.
- Lọc theo bậc: Loại bỏ nút có bậc nhỏ hơn 2 để giảm nhiễu.
- Chuẩn hóa chỉ số: Tính toán và chuẩn hóa các đặc trưng cấu trúc như bậc nút, số láng giềng chung, Jaccard, Adamic–Adar và Resource Allocation để sử dụng cho các mô hình học máy.
- Kiểm định chất lượng: Đánh giá lại các chỉ số mạng (bậc trung bình, hệ số gom cụm, đường kính mạng) sau khi làm sạch để đảm bảo cấu trúc tổng thể không bị biến dạng so với dữ liệu ban đầu.

D. Chiến lược lấy mẫu âm

Để cân bằng dữ liệu, các cạnh âm được chọn theo:

- Khoảng cách: Ưu tiên chọn các cặp nút có khoảng cách ngắn (2–3 bước) để đảm bảo khả năng hình thành cạnh là hợp lý.
- Theo bậc: Giữ phân bố bậc của các nút trong tập cạnh âm tương đồng với cạnh dương, tránh thiên lệch theo độ phổ biến.
- Theo cộng đồng: Kết hợp thông tin cộng đồng để đảm bảo cân bằng giữa các cặp nằm trong cùng cộng đồng và khác cộng đồng, giúp mô hình học được cả liên kết nội và ngoại nhóm.
- Ngẫu nhiên có kiểm soát: Việc chọn cạnh âm được thực hiện ngẫu nhiên nhưng tuân thủ các tiêu chí trên, đảm bảo dữ liệu huấn luyện đa dạng và ổn định.

E. Đảm bảo chất lượng dữ liệu

Sau khi hoàn tất tiền xử lý và lấy mẫu, dữ liệu được kiểm định lại để đảm bảo tính đầy đủ và độ tin cậy trước khi tiến hành huấn luyện mô hình

- Kiểm định thống kê: So sánh lại phân bố bậc, số lượng cạnh, đường kính mạng và hệ số gom cụm trước và sau xử lý.
- Kiểm tra tính cân bằng: Xác nhận tỷ lệ cạnh dương – âm, cũng như phân bố các đặc trưng, không bị lệch đáng kể giữa các tập huấn luyện và kiểm thử.
- Đầy đủ dữ liệu: Đảm bảo không mất nút hoặc cạnh quan trọng trong quá trình lọc và tách dữ liệu.

Bộ dữ liệu cuối cùng thu được phản ánh chính xác cấu trúc mạng cộng tác, giữ nguyên đặc trưng phân bố của mạng gốc và tạo nền tảng vững chắc cho các bước tách cạnh, huấn luyện mô hình và đánh giá hiệu suất trong phần thực nghiệm.

V. Thực nghiệm đánh giá

Trong phần này, chúng tôi tiến hành thiết kế thực nghiệm nhằm so sánh hiệu năng của các phương pháp dự đoán liên kết trên đồ thị. Hai kỹ thuật tách cạnh phổ biến được áp dụng là LCC Split và Shuffle Split, nhằm kiểm chứng sự ổn định của mô hình dưới các cách chia dữ liệu khác nhau. Các phương pháp được tổ chức thành bốn nhóm chính:

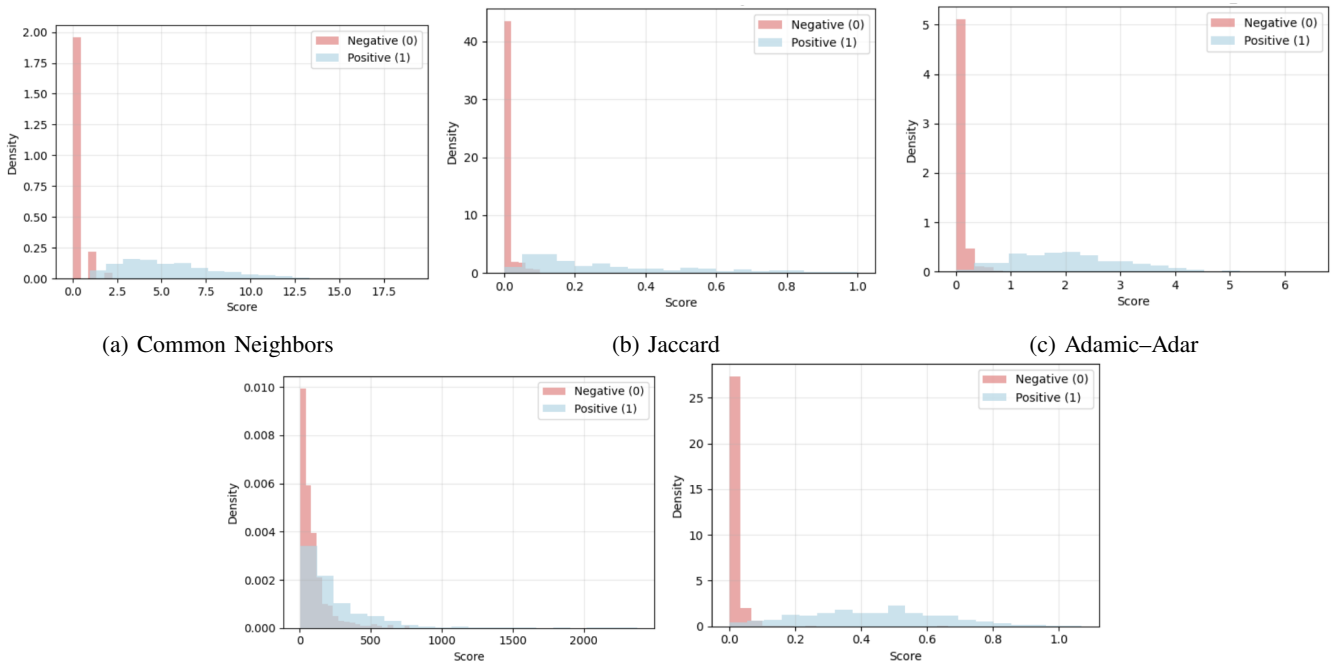
– Nhóm thứ nhất bao gồm các chỉ số heuristic truyền thống như Common Neighbors, Jaccard, Adamic–Adar, Resource Allocation và Preferential Attachment.
– Nhóm thứ hai dựa trên học máy với đặc trưng thủ công, trong đó các đặc trưng được xây dựng từ thông tin cấu trúc (bậc nút, số láng giềng chung, Jaccard, Adamic–Adar, Resource Allocation, Preferential Attachment và shortest path), sau đó được sử dụng để huấn luyện các mô hình Logistic Regression, Random Forest và Gradient Boosting.

– Nhóm thứ ba khai thác phương pháp biểu diễn học, sử dụng Node2Vec để sinh embedding cho các nút, tạo edge embedding bằng Hadamard product và tiến hành phân loại bằng Logistic Regression.
– Nhóm thứ tư áp dụng cách tiếp cận học sâu end-to-end với Graph Convolutional Network (GCN) hai tầng, trong đó biểu diễn nút được học trực tiếp từ ma trận kề và dự đoán liên kết được thực hiện thông qua dot-product scoring.

Tất cả các mô hình sẽ được chúng tôi đánh giá dựa trên các chỉ số AUC-ROC, AUC-PR và Average Precision (AP) trên cả hai phương pháp tách cạnh, nhằm cung cấp một so sánh toàn diện và làm rõ ưu nhược điểm của từng phương pháp.

A. *LCC Split*

Trước hết, chúng tôi tiến hành thực nghiệm với nhóm các phương pháp dựa trên chỉ số heuristic truyền thống.



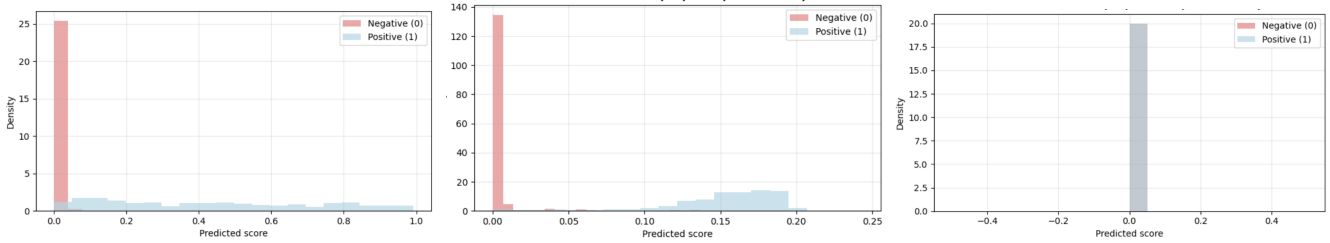
(a) Common Neighbors (b) Jaccard (c) Adamic–Adar

(d) Preferential Attachment (e) Resource Allocation

Hình 2: Phân phối điểm dự đoán của các phương pháp heuristic dưới LCC Split.

Qua Hình 2b và Hình 2e có thể thấy các chỉ số Jaccard và Resource Allocation thể hiện sự phân tách rõ rệt giữa cạnh dương và cạnh âm, trong đó các cạnh âm tập trung chủ yếu gần giá trị 0 còn các cạnh dương trải rộng hơn ở vùng giá trị cao. Trong khi đó, Hình 2d cho thấy chỉ số Preferential Attachment có sự chồng lấn lớn giữa hai lớp, do đó kém hiệu quả trong việc phân biệt. Các phân bố của Adamic–Adar (Hình 2c) và Common Neighbors (Hình 2a) cho thấy mức độ phân tách trung bình, với hiện tượng chồng lấn đáng kể ở vùng giá trị thấp. Nhìn chung, kết quả được biểu diễn qua Hình 2 cho thấy trong nhóm heuristic, Jaccard và Resource Allocation là hai phương pháp có tiềm năng đạt hiệu năng tốt khi đánh giá bằng các thước đo định lượng.

Sau đó chúng tôi tiếp tục tiến hành với các phương pháp ở nhóm 2, nơi các đặc trưng thủ công được kết hợp với các mô hình học máy cổ điển nhằm kiểm chứng khả năng phân biệt cạnh dương và cạnh âm.

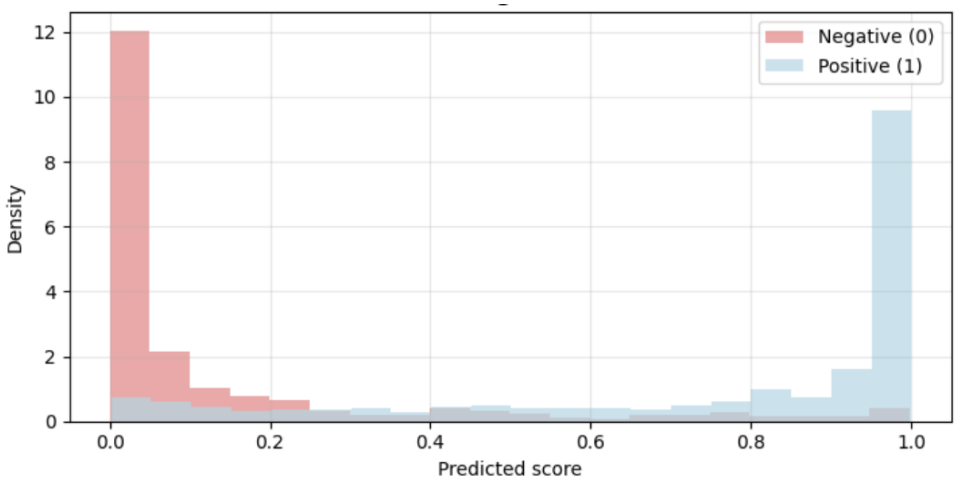


(a) Logistic Regression (b) Random Forest (c) Gradient Boosting

Hình 3: Phân phối điểm dự đoán của các mô hình học máy với đặc trưng thủ công khi áp dụng LCC Split.

Qua Hình 3a có thể thấy Logistic Regression cho phân phối cạnh âm tập trung sát giá trị 0, trong khi cạnh dương trải dài hơn về phía giá trị cao, cho thấy khả năng tách lớp tương đối tốt. Trong Hình 3b, Random Forest thể hiện sự phân tách rõ ràng hơn, khi cạnh âm gần như dồn hẳn về 0 và cạnh dương tập trung ở một vùng giá trị cao hơn (khoảng 0.1–0.2), chứng tỏ mô hình này khai thác đặc trưng thủ công hiệu quả nhất. Ngược lại, Hình 3c cho thấy Gradient Boosting không phân biệt được hai lớp, khi toàn bộ điểm dự đoán co cụm quanh 0, cho thấy mô hình này không phù hợp với tập đặc trưng hiện tại. Tổng thể, Random Forest nổi bật nhất trong nhóm, Logistic Regression ở mức trung bình, còn Gradient Boosting gần như thất bại trong việc phân biệt cạnh dương và cạnh âm. Tiếp theo, để kiểm chứng khả năng học biểu diễn từ cấu trúc đồ thị thay vì dựa vào đặc trưng thủ công, chúng tôi tiến hành thực nghiệm với mô hình ở nhóm 3, trong đó sử dụng Node2Vec để sinh embedding cho nút và Logistic Regression để phân loại cạnh.

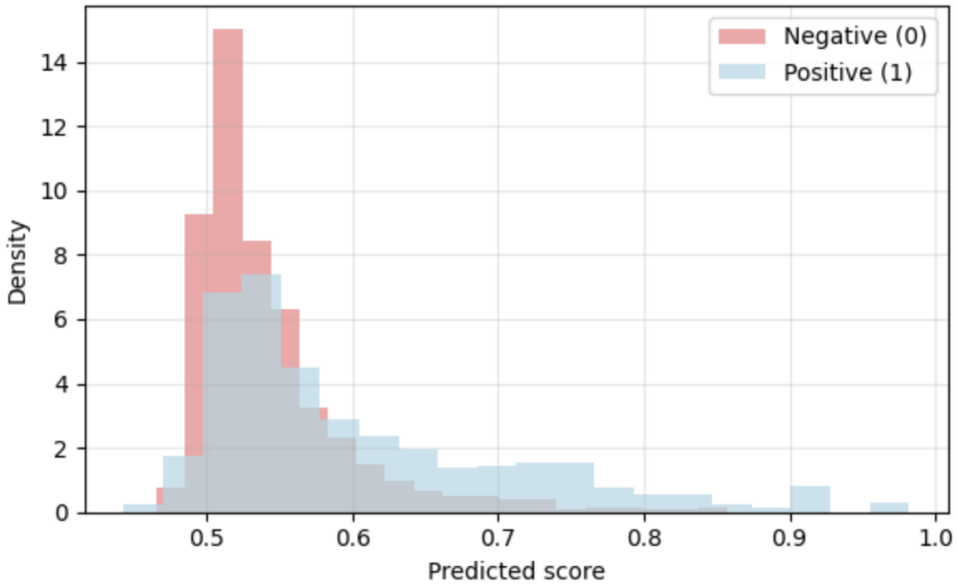
Hiệu suất của mô hình Node2Vec kết hợp Logistic Regression được biểu diễn trong Hình 4. Kết quả cho thấy các cạnh âm tập trung chủ yếu ở vùng giá trị gần 0, trong khi các cạnh dương phân bố rõ rệt quanh giá trị gần 1. Điều này chứng tỏ biểu diễn học từ Node2Vec đã mã hóa tốt thông tin cấu trúc của đồ thị, giúp bộ phân loại tuyến tính phân biệt cạnh dương và cạnh âm một cách hiệu quả hơn so với các phương pháp dựa trên đặc trưng thủ công.



Hình 4: Phân phối điểm dự đoán của mô hình Node2Vec kết hợp Logistic Regression khi áp dụng LCC Split.

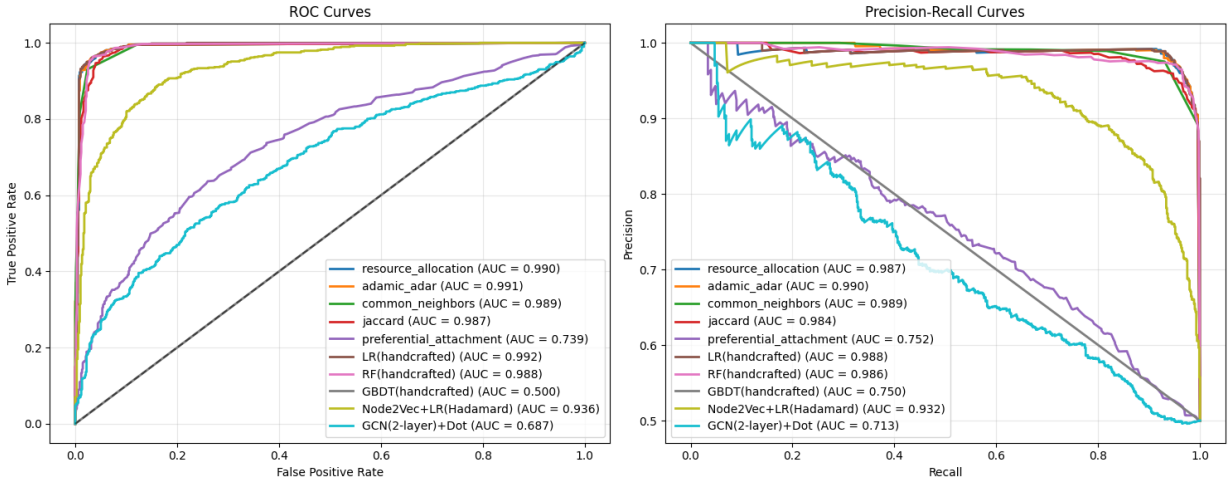
Như vậy, Node2Vec kết hợp Logistic Regression đã cho thấy hiệu quả tương đối rõ rệt nhờ khả năng học biểu diễn từ cấu trúc đồ thị. Tuy nhiên, phương pháp này vẫn cần một bước trung gian để tách quá trình học embedding và quá trình phân loại. Để đánh giá khả năng học end-to-end trực tiếp trên đồ thị, chúng tôi tiếp tục xem xét nhóm phương pháp cuối cùng, trong đó áp dụng Graph Convolutional Network (GCN) hai tầng để học biểu diễn nút và dự đoán liên kết thông qua dot-product scoring.

Hiệu suất của mô hình GCN hai tầng được biểu diễn trong Hình 5. Kết quả cho thấy phân phối điểm giữa hai lớp có sự chồng lấn đáng kể, khi cả cạnh dương và cạnh âm đều phân bố khá gần nhau quanh giá trị trung bình. Điều này cho thấy mô hình GCN trong thiết lập hiện tại chưa khai thác hiệu quả cấu trúc đồ thị để học biểu diễn nút phục vụ dự đoán liên kết. Nguyên nhân có thể đến từ việc mô hình còn đơn giản, chỉ sử dụng hai tầng tích chập và không có đặc trưng đầu vào giàu thông tin, dẫn đến hiện tượng underfitting. So với các nhóm phương pháp trước, hiệu suất của GCN thấp hơn rõ rệt, đặc biệt khi so sánh với Random Forest (nhóm 2) và Node2Vec (nhóm 3).



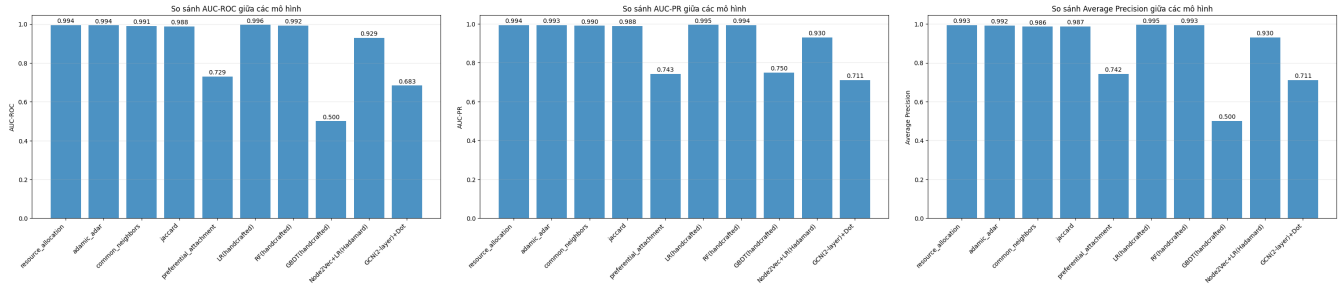
Hình 5: Phân phối điểm dự đoán của mô hình GCN hai tầng khi áp dụng LCC Split.

Để có một cái nhìn toàn diện về hiệu quả của các phương pháp ở kỹ thuật tách cạnh thứ nhất, chúng tôi trực quan hóa kết quả thông qua đường cong ROC và Precision–Recall, được trình bày trong Hình 6. Kết quả cho thấy Random Forest, Logistic Regression và các chỉ số heuristic như Resource Allocation và Adamic–Adar đạt hiệu quả nổi bật với diện tích dưới đường cong gần như tiệm cận 1, trong khi GCN và Gradient Boosting thể hiện rõ sự hạn chế khi các đường cong nằm thấp hơn đáng kể. Xu hướng trên hai loại đường cong là nhất quán, cho thấy sự ổn định của các mô hình mạnh, đồng thời củng cố các quan sát trước đó từ phân phối điểm của từng nhóm.



Hình 6: Đường cong ROC và Precision–Recall của tất cả mô hình khi áp dụng LCC Split.

Ngoài ra, để trực quan hơn về hiệu suất, chúng tôi cũng biểu diễn kết quả so sánh trên từng chỉ số AUC-ROC, AUC-PR và Average Precision như trong Hình 7.



(a) AUC-ROC (b) AUC-PR (c) Average Precision

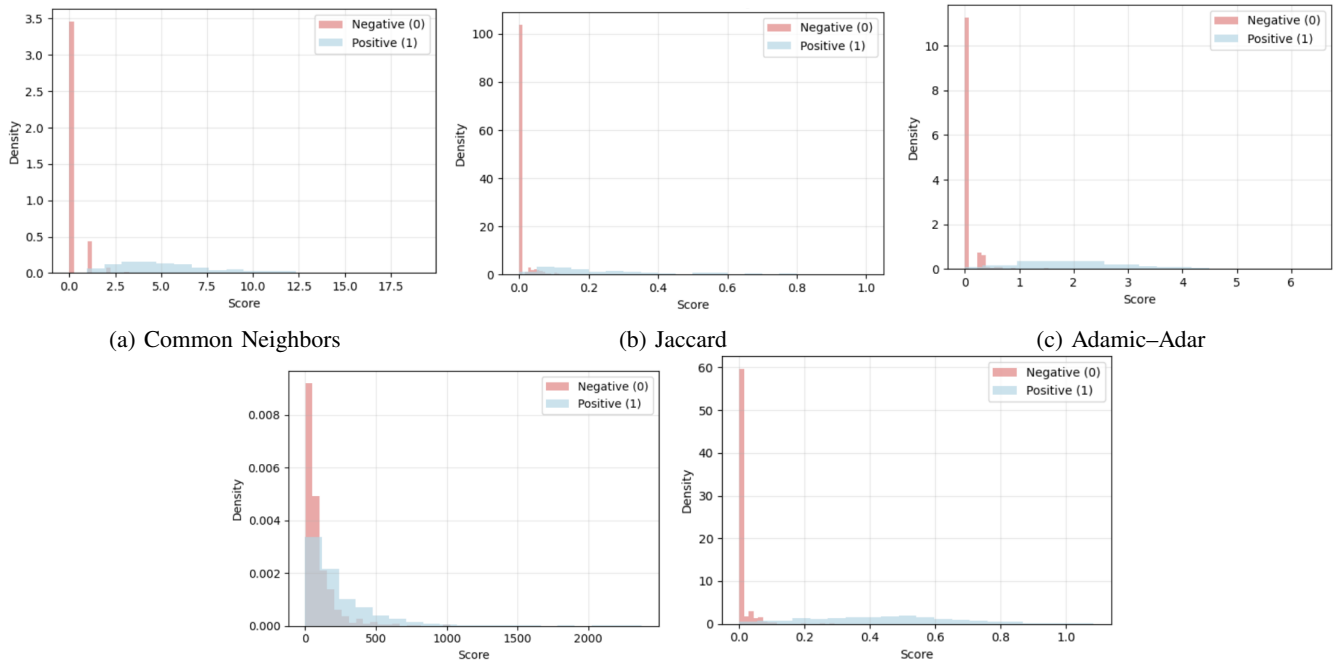
Hình 7: So sánh hiệu quả của các mô hình theo ba thước đo AUC-ROC, AUC-PR và Average Precision khi áp dụng LCC Split.

Qua Hình 7 có thể thấy xu hướng so sánh giữa các mô hình hoàn toàn nhất quán trên cả ba thước đo AUC-ROC, AUC-PR và Average Precision. Các chỉ số heuristic như Resource Allocation, Adamic–Adar và Common Neighbors đều đạt kết quả tiệm cận 1, cho thấy khả năng phân biệt vượt trội và ổn định. Random Forest cũng giữ vị trí hàng đầu, khẳng định tính hiệu quả của việc kết hợp đặc trưng thủ công với mô hình cây. Logistic Regression và Jaccard tuy có giá trị thấp hơn đôi chút nhưng vẫn duy trì ở mức rất cao, thể hiện tính đáng tin cậy. Ngược lại, Gradient Boosting và GCN hai tầng thể hiện rõ sự hạn chế với kết quả thấp hơn đáng kể, trong khi Node2Vec kết hợp Logistic Regression nằm ở mức trung bình. Nhìn chung, kết quả này củng cố các quan sát trước đó từ phân phối điểm và đường cong ROC/PR, đồng thời đưa ra một so sánh định lượng rõ ràng hơn về ưu nhược điểm của từng nhóm phương pháp.

Tiếp theo để kiểm chứng một cách tổng quan, chúng tôi tiếp tục tiến hành thực nghiệm với kỹ thuật tách cạnh Shuffle Split.

B. Shuffle Split

Tương tự như ở kỹ thuật tách cạnh thứ nhất, ở kỹ thuật tách cạnh này chúng tôi cũng tiến hành đánh giá lần lượt trên bốn nhóm phương pháp đã nêu. Đầu tiên là nhóm các phương pháp heuristic, trong đó bao gồm Common Neighbors, Jaccard, Adamic–Adar, Preferential Attachment và Resource Allocation.



(a) Common Neighbors (b) Jaccard (c) Adamic–Adar

(d) Preferential Attachment (e) Resource Allocation

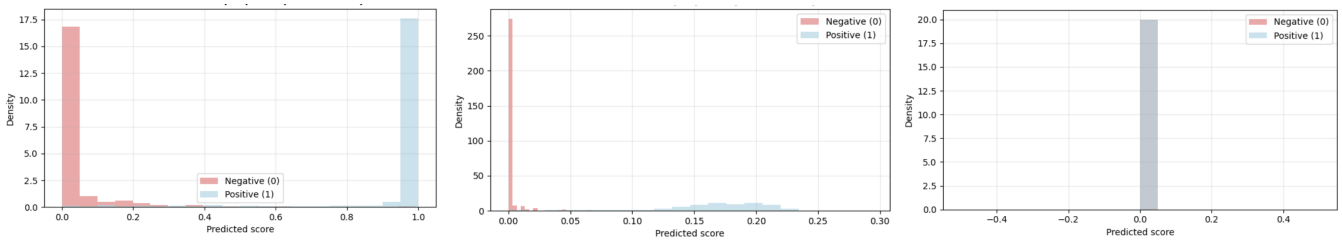
Hình 8: Phân phối điểm dự đoán của các mô hình heuristics khi áp dụng Shuffle Split.

Qua Hình 8 có thể thấy các chỉ số heuristic nhìn chung vẫn duy trì được xu hướng quan sát từ kỹ thuật tách cạnh LCC Split. Cụ thể, Jaccard và Resource Allocation tiếp tục thể hiện sự phân tách khá rõ ràng giữa cạnh dương và cạnh âm, với phần

lớn cạnh âm tập trung gần giá trị thấp và cạnh dương trải rộng hơn ở các giá trị cao. Adamic–Adar và Common Neighbors cho thấy mức độ phân biệt ở mức trung bình, khi vẫn tồn tại sự chồng lấn giữa hai lớp ở vùng giá trị thấp. Trong khi đó, Preferential Attachment vẫn là chỉ số yếu nhất, với phân phối điểm của hai lớp hầu như chồng khít, gây khó khăn cho việc phân loại. Kết quả này cho thấy tính ổn định của các heuristics mạnh (Jaccard, Resource Allocation) trên cả hai kỹ thuật tách cạnh, đồng thời củng cố kết luận rằng Preferential Attachment không phải là một chỉ số hiệu quả cho bài toán dự đoán liên kết.

Tiếp theo, chúng tôi xem xét nhóm 2, bao gồm các mô hình học máy sử dụng đặc trưng thủ công. Các mô hình Logistic Regression, Random Forest và Gradient Boosting sẽ được đánh giá trên tập dữ liệu được chia theo Shuffle Split nhằm kiểm chứng mức độ ổn định và khả năng tổng quát của chúng.

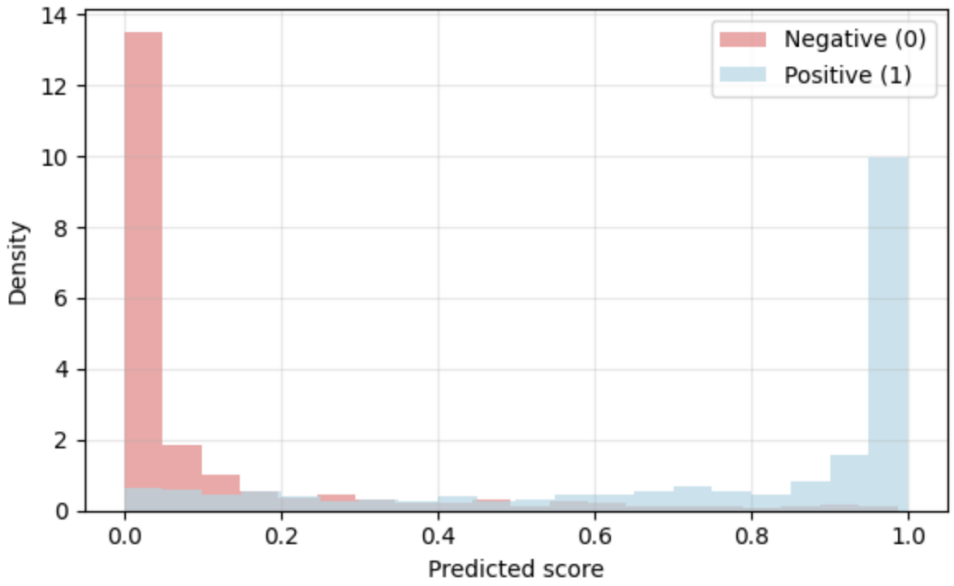
Các mô hình ở nhóm 2 được thể hiện rõ ở Hình 9.



(a) Logistic Regression (b) Random Forest (c) Gradient Boosting

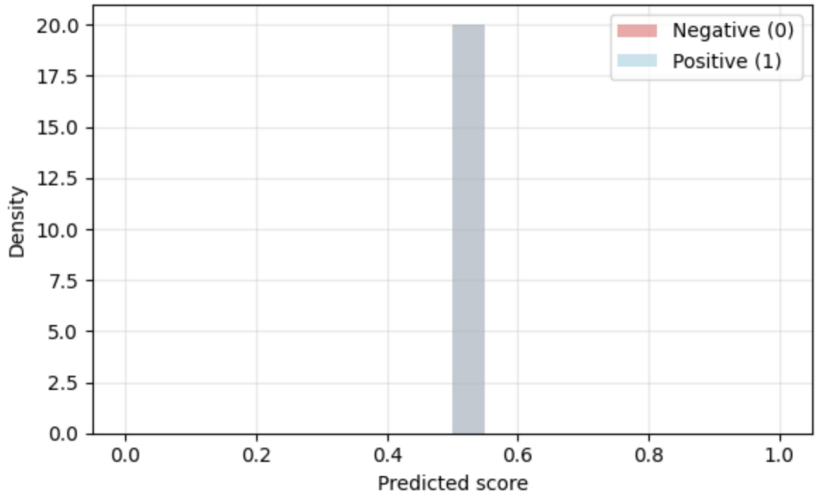
Hình 9: Phân phối điểm dự đoán của các mô hình học máy khi áp dụng Shuffle Split.

Qua đó có thể thấy Logistic Regression cho kết quả phân tách tương đối tốt, với đa số cạnh âm tập trung gần 0 và cạnh dương nghiêng hẳn về phía 1. Random Forest tiếp tục khẳng định hiệu quả nổi bật khi hai lớp được tách biệt rõ ràng, cho thấy tính ổn định ngay cả khi thay đổi kỹ thuật tách cạnh. Ngược lại, Gradient Boosting không thể hiện được khả năng phân loại khi hầu hết dự đoán đều co cụm quanh giá trị gần 0, tương tự quan sát ở LCC Split. Điều này cho thấy Random Forest vẫn là mô hình mạnh nhất trong nhóm này, trong khi Logistic Regression đáng tin cậy và Gradient Boosting kém hiệu quả.



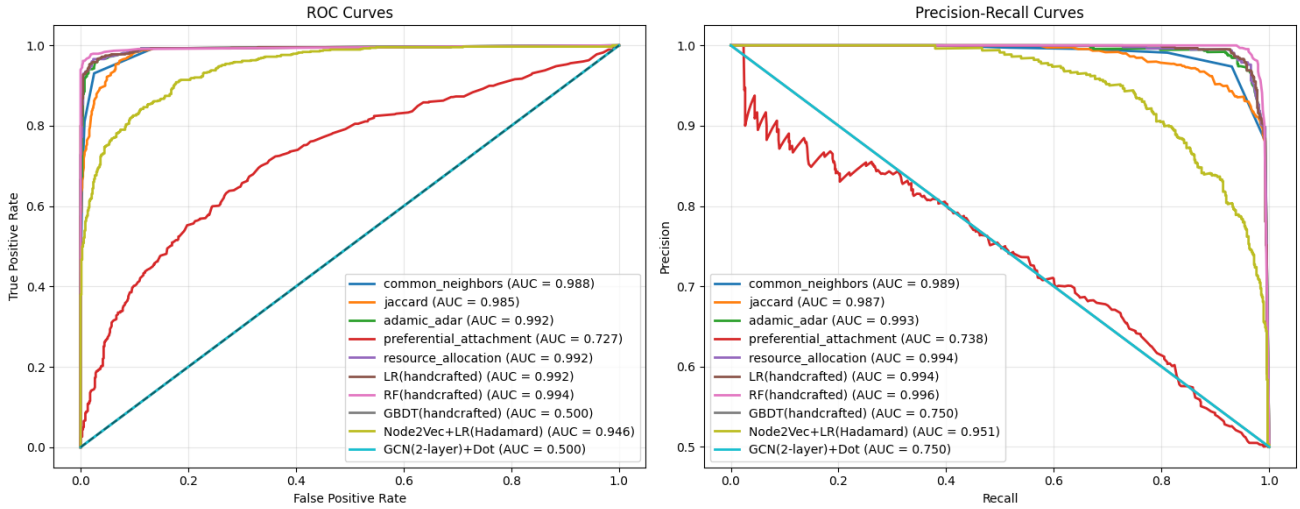
Hình 10: Phân phối điểm dự đoán của mô hình Node2Vec kết hợp Logistic Regression khi áp dụng Shuffle Split.

Các mô hình ở nhóm 3 được thể hiện rõ ở Hình 10. Kết quả cho thấy phương pháp Node2Vec kết hợp Logistic Regression tiếp tục duy trì khả năng phân biệt hai lớp cạnh khá rõ ràng. Cạnh âm chủ yếu phân bố ở gần 0, trong khi cạnh dương tập trung ở gần 1, tạo nên sự tách biệt rõ rệt. So với hai nhóm trước, kết quả này khẳng định ưu thế của cách tiếp cận embedding học từ cấu trúc đồ thị, giúp mô hình đạt hiệu quả ổn định ngay cả khi thay đổi kỹ thuật tách cạnh. Tuy nhiên, do vẫn cần một bước trung gian giữa học embedding và phân loại, hiệu quả chưa hoàn toàn vượt trội so với Random Forest hoặc các heuristic mạnh như Resource Allocation.



Hình 11: Phân phối điểm dự đoán của mô hình GCN hai tầng khi áp dụng Shuffle Split.

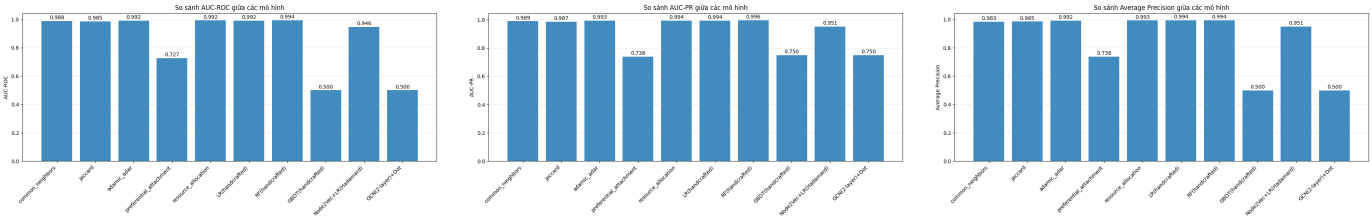
Các mô hình ở nhóm 4 được thể hiện rõ ở Hình 11. Kết quả cho thấy mô hình GCN hai tầng không phân biệt được hai lớp cạnh, khi toàn bộ giá trị dự đoán đều co cụm quanh một điểm duy nhất (xấp xỉ 0.5). Điều này dẫn tới sự chồng lấn hoàn toàn giữa cạnh dương và cạnh âm, làm cho mô hình không thể hiện được năng lực học biểu diễn trong bối cảnh này. So với các phương pháp đã trình bày ở nhóm trước, đặc biệt là Random Forest hay Node2Vec, GCN cho thấy hiệu quả kém vượt trội, phản ánh hạn chế của việc huấn luyện end-to-end trên đồ thị với quy mô và dữ liệu hiện tại.



Hình 12: Đường cong ROC và Precision–Recall của tất cả mô hình khi áp dụng Shuffle Split.

Để có một cái nhìn toàn diện hơn, chúng tôi trực quan hoá kết quả của tất cả mô hình thông qua đường cong ROC và Precision–Recall, được thể hiện ở Hình 12. Kết quả cho thấy các heuristics mạnh như Resource Allocation, Adamic–Adar và Common Neighbors đạt hiệu quả gần như hoàn hảo với diện tích dưới đường cong tiệm cận 1.0. Random Forest tiếp tục thể hiện vượt trội và ổn định, trong khi Logistic Regression và Jaccard cũng duy trì được hiệu quả cao. Ngược lại, Gradient Boosting và GCN hai tầng có hiệu suất thấp, với AUC dao động quanh mức 0.5–0.75, cho thấy khả năng phân biệt hạn chế. Node2Vec kết hợp Logistic Regression cho kết quả khá tốt nhưng chưa đạt mức của Random Forest hay các heuristics hàng đầu. Tổng thể, kết quả ở Shuffle Split củng cố quan sát từ LCC Split, khẳng định tính ổn định của các phương pháp heuristic và Random Forest, đồng thời chỉ ra hạn chế của GCN trong bối cảnh này.

Tiếp theo nhằm củng cố những quan sát từ đường cong ROC và Precision–Recall, chúng tôi tiếp tục trình bày so sánh định lượng hiệu suất của các mô hình theo ba thước đo phổ biến: AUC-ROC, AUC-PR và Average Precision. Kết quả trực quan hoá được thể hiện ở Hình 13.



(a) AUC-ROC (b) AUC-PR (c) Average Precision

Hình 13: So sánh hiệu suất của các mô hình theo ba thước đo AUC-ROC, AUC-PR, Average Precision khi áp dụng Shuffle Split.

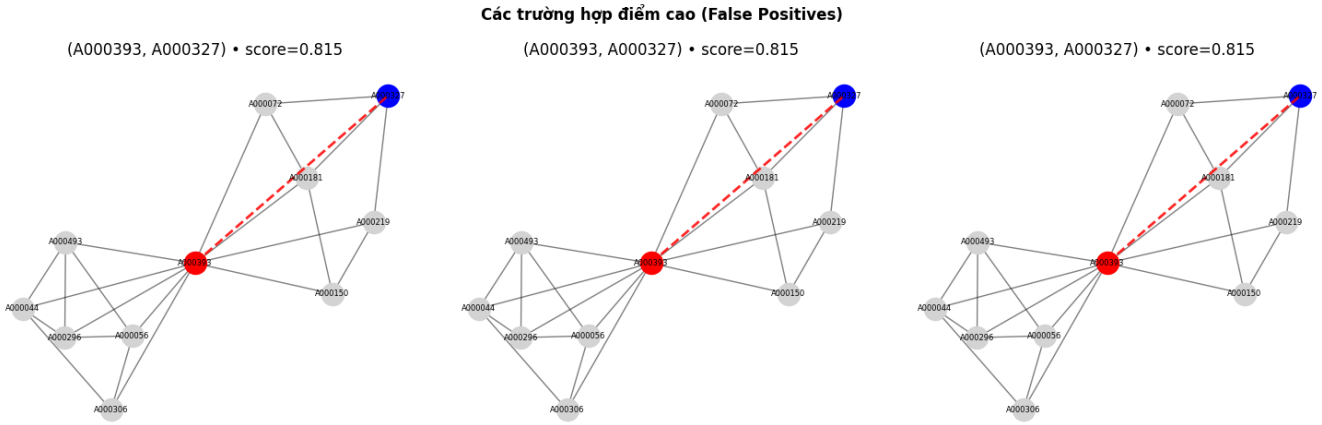
Kết quả cho thấy các heuristics như Resource Allocation, Adamic–Adar và Common Neighbors đạt điểm số gần như tuyệt đối trên cả ba thước đo, phản ánh tính ổn định và hiệu quả cao. Random Forest tiếp tục nổi bật trong nhóm phương pháp học máy với đặc trưng thủ công, trong khi Logistic Regression cũng giữ mức ổn định nhưng thấp hơn một chút. Ngược lại, Gradient Boosting và GCN hai tầng cho kết quả thấp nhất, thể hiện hạn chế trong việc khai thác thông tin cấu trúc. Node2Vec kết hợp Logistic Regression mang lại kết quả khá nhưng vẫn chưa bằng Random Forest hay các heuristics mạnh. Những so sánh này nhất quán với phân tích trực quan từ đường cong ROC và PR, đồng thời củng cố nhận định về ưu thế của heuristics và Random Forest trong bối cảnh Shuffle Split.

Bảng III: Kết quả so sánh hiệu quả các mô hình dưới hai kỹ thuật tách cạnh (làm tròn 3 chữ số).

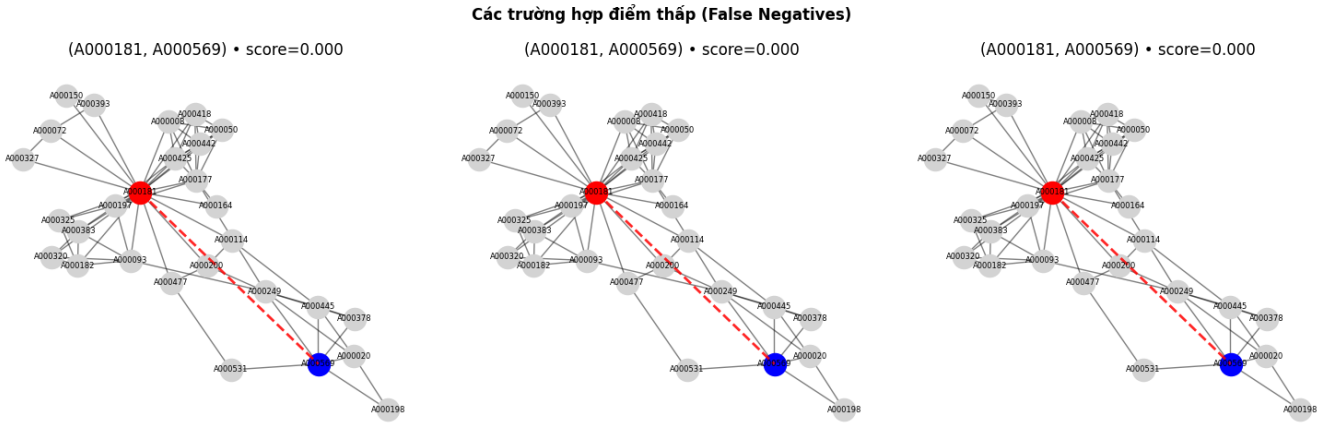
Kỹ thuật tách cạnh	Mô hình	AUC-ROC	AUC-PR	AP
LCC Split	LR (handcrafted)	0.996	0.995	0.995
	RF (handcrafted)	0.992	0.994	0.993
	Resource Allocation	0.994	0.994	0.993
	Adamic–Adar	0.994	0.993	0.992
	Common Neighbors	0.991	0.990	0.986
	Jaccard	0.988	0.988	0.987
	Node2Vec + LR (Hadamard)	0.929	0.930	0.930
	GBDT (handcrafted)	0.500	0.750	0.500
	Preferential Attachment	0.729	0.743	0.742
	GCN (2-layer) + Dot	0.683	0.711	0.711
Shuffle Split	RF (handcrafted)	0.994	0.996	0.994
	Resource Allocation	0.992	0.994	0.993
	LR (handcrafted)	0.992	0.994	0.994
	Adamic–Adar	0.992	0.993	0.992
	Common Neighbors	0.988	0.989	0.983
	Jaccard	0.985	0.987	0.985
	Node2Vec + LR (Hadamard)	0.946	0.951	0.951
	GBDT (handcrafted)	0.500	0.750	0.500
	GCN (2-layer) + Dot	0.500	0.750	0.500
	Preferential Attachment	0.727	0.738	0.738

Thông qua kết quả ở Bảng III cho thấy các chỉ số heuristic (Resource Allocation, Adamic–Adar, Common Neighbors, Jaccard) và các mô hình học máy dựa trên đặc trưng thủ công đều đạt hiệu suất rất cao và ổn định. Cụ thể, Logistic Regression với đặc trưng thủ công đạt kết quả tốt nhất trong LCC Split (AUC-ROC 0.996, AUC-PR 0.995, AP 0.995), trong khi Random Forest nổi bật nhất ở Shuffle Split với AUC-ROC 0.994 và AUC-PR 0.996. Các chỉ số heuristic như Resource Allocation và Adamic–Adar cũng đạt kết quả gần như tương đương, xác nhận sự ổn định và sức mạnh của các phương pháp dựa trên cấu trúc cục bộ của đồ thị. Ngược lại, các mô hình học sâu end-to-end như GCN hai tầng và Gradient Boosting cho thấy hiệu suất thấp đáng kể, thậm chí gần mức ngẫu nhiên trong một số trường hợp, phản ánh hạn chế khi thiếu đặc trưng đầu vào phù hợp hoặc cấu trúc mô hình chưa đủ mạnh. Phương pháp Node2Vec kết hợp Logistic Regression đạt kết quả trung bình (AUC-ROC khoảng 0.93–0.95), cho thấy tiềm năng của học biểu diễn nhưng chưa vượt qua được heuristics và Random Forest.

Kết quả này cho thấy các phương pháp đơn giản hơn như chỉ số heuristic hoặc Logistic Regression/Random Forest với đặc trưng thủ công vẫn duy trì ưu thế rõ rệt so với các tiếp cận phức tạp hơn. Trong trường hợp này, do Logistic Regression với đặc trưng thủ công đạt hiệu suất cao nhất và ổn định trên cả hai kỹ thuật tách cạnh, chúng tôi lựa chọn mô hình này làm đại diện cho phần phân tích chi tiết các trường hợp dự đoán trong phần tiếp theo.



Hình 14: Các trường hợp điểm điểm thấp



Hình 15: Các trường hợp điểm cao

Quan sát Hình 14 và Hình 15 cho thấy một số trường hợp mô hình Logistic Regression (handcrafted features) dự đoán sai. False Positives thường xuất hiện khi hai nút chia sẻ nhiều láng giềng chung nên được mô hình gán xác suất cao mặc dù thực tế không có cạnh. Ngược lại, False Negatives xuất hiện ở các cạnh thật nhưng hai nút nằm xa nhau trong đồ thị hoặc kết nối qua trung gian yếu, dẫn tới điểm dự đoán rất thấp.

VI. THẢO LUẬN

Trong phần này, chúng tôi phân tích và diễn giải các kết quả thực nghiệm, làm rõ ý nghĩa của những quan sát định lượng, chỉ ra tiềm năng mở rộng của nghiên cứu trong tương lai.

A. Các phát hiện nghiên cứu chính

Thực nghiệm trên hai kỹ thuật tách cạnh LCC Split và Shuffle Split , cho thấy các mô hình thể hiện xu hướng kết quả ổn định. Qua đó, có thể tổng hợp một số phát hiện nổi bật như sau:

- Ưu thế của các chỉ số heuristic: Các chỉ số Resource Allocation (RA), Adamic–Adar (AA) và Common Neighbors (CN) đạt hiệu suất rất cao, phần lớn thông tin cần thiết để dự đoán ai sẽ kết nối với ai nằm trong cấu trúc lân cận trực tiếp của họ. Chỉ số này tuân theo nguyên tắc "bạn của bạn là bạn của bạn", có nghĩa hai nút có nhiều láng giềng chung thì khả năng kết nối càng lớn, đúng trong các mạng xã hội hoặc mạng hợp tác có tính cộng đồng cao. Trong đó, RA gán trọng số cao hơn cho các láng giềng chung có bậc nhỏ, làm nổi bật các kết nối yếu nhưng mang giá trị cấu trúc quan trọng; còn AA điều chỉnh trọng số theo logarit bậc nút, giúp giảm ảnh hưởng của các "ngôi sao mạng" và duy trì sự cân bằng khi dự đoán liên kết. Ngược lại, chỉ số Preferential Attachment (PA) cho kết quả thấp hơn đáng kể (AUC khoảng 0.72–0.74) do chỉ dựa vào tích bậc của hai nút, phản ánh cơ chế "giàu càng giàu". Cách tiếp cận này phù hợp với mạng có phân bố bậc lũy thừa rõ rệt nhưng kém hiệu quả trong dữ liệu hiện tại, cấu trúc mạng không tuân theo quy luật đó. Các chỉ số heuristic dựa trên láng giềng chung vẫn cho thấy ưu thế rõ rệt và là công cụ dự đoán đáng tin cậy nhất trong các mạng có cấu trúc cộng đồng rõ ràng.

- Tính ổn định của mô hình học máy cổ điển: Các mô hình Logistic Regression (LR) và Random Forest (RF) với đặc trưng thủ công đạt hiệu suất rất cao và ổn định trên hai kỹ thuật tách cạnh. Logistic Regression hoạt động hiệu quả trong LCC Split (AUC-ROC = 0.996) nhờ khả năng phân tách tuyến tính tốt giữa cạnh dương và âm, Random Forest đạt kết quả cao ở Shuffle Split (AUC-PR = 0.996) do khai thác được mối quan hệ phi tuyến giữa các đặc trưng cấu trúc. Kết quả cho thấy khi các đặc trưng được xây dựng phù hợp, những mô hình cổ điển vẫn có thể đạt kết quả vượt qua các mô hình học sâu phức tạp hơn. Gradient Boosting cho thấy hiệu quả thấp do tập đặc trưng hiện tại chưa phù hợp với cơ chế tăng cường tuần tự. Các mô hình học máy cổ điển vừa đảm bảo độ chính xác cao, dễ huấn luyện và giải thích, phù hợp với quy mô dữ liệu trong nghiên cứu này. Điều này xuất phát từ việc các đặc trưng đầu vào chủ yếu là các chỉ số cấu trúc đơn giản (Common Neighbors, Adamic–Adar, Resource Allocation, Jaccard), có mối quan hệ gần như tuyến tính. Trong trường hợp này, mô hình Boosting khó phát huy ưu thế về khả năng mô hình hóa các quan hệ phi tuyến phức tạp. Hơn nữa, quy mô dữ liệu nhỏ làm giảm khả năng tăng cường dần của Gradient Boosting, khiến mô hình dễ rơi vào tình trạng overfitting hoặc không cải thiện đáng kể so với Logistic Regression.

- Hạn chế của các mô hình phức tạp: Mô hình Graph Convolutional Network (GCN) hai tầng cho kết quả thấp nhất, với điểm dự đoán tập trung quanh 0.5 cho cả hai lớp. Nguyên nhân là do mô hình chỉ lan truyền thông tin trong phạm vi láng giềng gần nhất, thiếu đặc trưng đầu vào phong phú. Dữ liệu tĩnh và nhỏ, GCN không thể học được đặc trưng biểu diễn có ý nghĩa, có hiện tượng underfitting. Các mô hình học sâu phát huy hiệu quả khi được áp dụng cho mạng lớn, có nhiều thuộc tính nút hoặc yếu tố thời gian. Ngoài ra, do mạng hợp tác chỉ bao gồm một loại nút (diễn viên) và không có thuộc tính bổ sung như thể loại phim hay vai diễn, GCN gặp khó khăn trong việc học các mối quan hệ ngữ nghĩa phức tạp. Khi số lượng láng giềng ít, tín hiệu lan truyền bị suy giảm nhanh qua các lớp tích chập, gây ra hiện tượng over-smoothing khiến các nút có biểu diễn gần giống nhau. Kết quả là mô hình không phân biệt được giữa cặp có và không có liên kết, thể hiện rõ qua việc điểm dự đoán tập trung quanh 0.5.

- Vai trò của học biểu diễn: Phương pháp Node2Vec kết hợp Logistic Regression đạt hiệu suất tốt (AUC-ROC khoảng 0.93–0.95), khả năng học được biểu diễn tiềm ẩn từ cấu trúc đồ thị. So với các mô hình dựa trên đặc trưng thủ công, Node2Vec khai thác được mối quan hệ xa hơn giữa các nút thông qua cơ chế random walk, giúp phát hiện các kết nối tiềm năng vượt ra ngoài láng giềng trực tiếp. Do quá trình học embedding và phân loại được thực hiện tách biệt, mô hình chưa tận dụng được lợi thế tối ưu hóa end-to-end, Node2Vec vẫn cho thấy tiềm năng cao và là nền tảng tốt để mở rộng sang các mô hình biểu diễn phức tạp hơn như GraphSAGE hoặc GAT trong tương lai.

- Kết luận chung : Các phương pháp đơn giản dựa trên cấu trúc cục bộ và học máy cổ điển duy trì ưu thế vượt trội so với các mô hình phức tạp hơn. Các chỉ số heuristic và Logistic Regression/Random Forest đạt hiệu năng cao, ổn định và dễ triển khai thực tế. Các mô hình học sâu như GCN hay Gradient Boosting chưa phù hợp với quy mô dữ liệu nhỏ và thiếu đặc trưng đầu vào. Node2Vec thể hiện tiềm năng đáng kể trong việc học biểu diễn cấu trúc mạng. Trong tương lai, việc mở rộng quy mô dữ liệu, bổ sung đặc trưng nội dung và thử nghiệm các biến thể của mạng học sâu như GraphSAGE hoặc GAT có thể giúp khắc phục những hạn chế hiện tại và nâng cao khả năng khái quát hóa của mô hình.

B. Gợi ý thực tiễn

Những phát hiện trên mang lại một số gợi ý thực tiễn quan trọng:

- Phân tích cấu trúc mạng xã hội: Các chỉ số heuristic như Resource Allocation (RA) và Adamic–Adar (AA) có thể được áp dụng trực tiếp để xác định các cặp nút tiềm năng có khả năng hình thành liên kết mới trong tương lai. Trong mạng xã hội hoặc mạng cộng tác, hai cá nhân có nhiều bạn chung hoặc có "láng giềng cấu trúc" gần nhau thường có xu hướng kết nối. Các chỉ số này là công cụ hữu hiệu trong việc phát hiện quan hệ tiềm năng và dự báo mối quan hệ sắp hình thành, giúp hỗ trợ xây dựng các mô hình khuyến nghị kết nối (link recommendation) hoặc gợi ý hợp tác (collaboration suggestion).

- Ứng dụng trong hệ thống gợi ý: Các mô hình học máy cổ điển như Logistic Regression (LR) và Random Forest (RF), khi được huấn luyện trên các đặc trưng cấu trúc, có thể được tích hợp hiệu quả vào hệ thống gợi ý kết nối người dùng. So với các mô hình học sâu phức tạp, dễ triển khai hơn, dễ cập nhật, diễn giải rõ ràng, và vẫn đạt độ chính xác rất cao, phù hợp cho các ứng dụng yêu cầu phản hồi nhanh và ổn định.

- Giám sát và phát hiện bất thường: Các trường hợp False Positive (dự đoán có cạnh nhưng thực tế không có) và False Negative (bỏ sót cạnh thực tế) quan sát trong kết quả có thể được khai thác để phát hiện kết nối bất thường hoặc hành vi tiềm ẩn. Một cặp nút có điểm dự đoán cao nhưng không tồn tại liên kết thực tế có thể là mối quan hệ sắp hình thành hoặc liên kết đáng ngờ, hỗ trợ tốt cho bài toán phát hiện gian lận hoặc giám sát mạng giao dịch.

- Xây dựng mô hình lai: Kết quả cho thấy việc kết hợp các chỉ số heuristic với học máy là một hướng đi tiềm năng. Các chỉ số như RA hoặc AA có thể được sử dụng làm đặc trưng đầu vào cho Logistic Regression hoặc Random Forest, mô hình vừa giữ được khả năng diễn giải và tốc độ của heuristic, vừa tận dụng được sức mạnh tổng quát của học máy. Cách tiếp cận này giúp cân bằng giữa độ chính xác, tính ổn định và khả năng mở rộng trong ứng dụng thực tế.

C. Ý nghĩa và nhận xét

Kết quả thực nghiệm cho thấy các phương pháp dựa trên cấu trúc mạng cục bộ là hướng tiếp cận hiệu quả nhất cho bài toán dự đoán liên kết. Các chỉ số như Common Neighbors, Resource Allocation hay Adamic–Adar tận dụng tốt đặc tính

"homophily" trong mạng xã hội, các nút có đặc điểm tương đồng thường có xu hướng kết nối với nhau. Các mô hình học sâu như GCN chưa phát huy hiệu quả do dữ liệu hiện tại là đồ thị tĩnh, không có yếu tố thời gian và thiếu đặc trưng nút, mô hình không học được hình thành cạnh. Phương pháp Node2Vec cho thấy tiềm năng đáng kể nhờ khả năng học biểu diễn phi tuyến và khái quát tốt hơn. Các mô hình đơn giản dựa trên cấu trúc vẫn là lựa chọn tối ưu trong bối cảnh dữ liệu vừa và nhỏ, là nền tảng quan trọng để phát triển các mô hình phức tạp hơn trong tương lai.

D. Điểm mạnh và hạn chế của nghiên cứu

Điểm mạnh:

- Thực hiện so sánh toàn diện giữa nhiều nhóm phương pháp, bao gồm heuristic, học máy cổ điển, học biểu diễn và học sâu, giúp đánh giá được ưu nhược điểm của từng hướng tiếp cận.
- Phân tích trực quan thông qua biểu đồ, phân phối điểm, và đường cong ROC/PR, giúp người đọc dễ dàng quan sát và hiểu rõ hiệu suất mô hình.
- Dữ liệu được xử lý kỹ, đảm bảo độ tin cậy của kết quả.
- Sử dụng hai kỹ thuật tách cạnh (LCC Split và Shuffle Split) giúp kiểm chứng được tính ổn định và khả năng tổng quát của các mô hình.

Hạn chế:

- Dữ liệu nhỏ, các mô hình học sâu (như GCN) chưa có đủ thông tin để học hiệu quả.
- Quá trình xử lý và làm sạch dữ liệu vẫn thực hiện thủ công, mất thời gian và dễ sai sót.
- Dữ liệu hiện tại là tĩnh, chưa phản ánh được sự thay đổi của mạng theo thời gian.
- Thiếu đặc trưng của nút, dẫn đến việc các mô hình học sâu chưa tận dụng được hết khả năng lan truyền thông tin trong đồ thị.

E. Định hướng nghiên cứu trong tương lai

Từ hạn chế đã nêu, một số hướng mở rộng có thể được xem xét như sau:

- Khắc phục hạn chế dữ liệu: Tập trung cải thiện chất lượng dữ liệu bằng cách chuẩn hóa, loại bỏ trùng lặp và xử lý sai lệch trong quá trình thu thập.
- Phân tích mạng động : Mở rộng nghiên cứu theo hướng phân tích và dự đoán sự thay đổi liên kết trong mạng ở các giai đoạn khác nhau nhằm nắm bắt yếu tố thời gian.
- Kết hợp nhiều phương pháp: Xây dựng mô hình lai giữa heuristic, học máy và học biểu diễn để tận dụng ưu điểm của từng hướng, giúp cải thiện độ chính xác và khả năng tổng quát hóa.
- Cải tiến mô hình học sâu: Thử nghiệm các kiến trúc tiên tiến hơn như GraphSAGE, Graph Attention Networks (GAT) hoặc Heterogeneous GNN để tăng khả năng học biểu diễn và khai thác quan hệ đa dạng trong đồ thị.
- Mở rộng phạm vi ứng dụng: Áp dụng và kiểm chứng mô hình trên các loại mạng khác như mạng xã hội, mạng học thuật hoặc mạng sản xuất để đánh giá tính tổng quát.

VII. KẾT LUẬN

Trong nghiên cứu này, chúng tôi đã tiến hành đánh giá và so sánh bốn nhóm phương pháp dự đoán liên kết trên mạng đồ thị, bao gồm: các chỉ số heuristic, mô hình học máy sử dụng đặc trưng thủ công, phương pháp học biểu diễn Node2Vec, và mô hình học sâu end-to-end GCN. Thực nghiệm trên hai kỹ thuật tách cạnh phổ biến là LCC Split và Shuffle Split, kiểm chứng độ ổn định và khả năng tổng quát của các mô hình dưới các cách chia dữ liệu khác nhau.

Kết quả cho thấy các chỉ số heuristic như Resource Allocation, Adamic–Adar và Common Neighbors đạt hiệu suất rất cao và ổn định, chỉ số AUC-ROC và AUC-PR gần tiệm cận 1.0. Các mô hình học máy truyền thống như Logistic Regression và Random Forest khi sử dụng đặc trưng thủ công cũng cho kết quả tương đương, trong đó Random Forest có khả năng tổng quát tốt nhất ở cả hai kỹ thuật tách cạnh. Các mô hình phức tạp hơn như Gradient Boosting và GCN hai tầng chưa đạt kết quả tốt, chủ yếu do đặc trưng dữ liệu hiện tại còn hạn chế: đồ thị tĩnh, chưa đủ lớn và thiếu thông tin thuộc tính của nút, khiến các mô hình học sâu chưa phát huy được.

Từ các kết quả trên, có thể thấy đối với dữ liệu đồ thị quy mô vừa và mang tính tĩnh, những phương pháp đơn giản dựa trên cấu trúc cục bộ vẫn mang lại độ chính xác cao, ổn định và dễ triển khai thực tế. Đây là hướng tiếp cận phù hợp cho các bài toán dự đoán liên kết trong mạng xã hội, mạng cộng tác hoặc hệ thống khuyến nghị.

Trong tương lai, nghiên cứu sẽ tập trung vào khắc phục các hạn chế về dữ liệu nhằm nâng cao chất lượng đầu vào cho mô hình. Cần đẩy mạnh việc chuẩn hóa, loại bỏ trùng lặp và phát hiện sai lệch trong quá trình thu thập dữ liệu,. Bên cạnh đó, mở rộng nghiên cứu sang mạng động theo thời gian sẽ giúp mô hình nắm bắt được sự thay đổi và tiến hóa của các liên kết, phản ánh chính xác hơn hành vi phát triển của mạng. Cuối cùng, hướng kết hợp giữa heuristic, học máy và học biểu diễn vẫn là tiềm năng, cho phép tận dụng ưu điểm của từng phương pháp nhằm đạt được hiệu quả dự đoán cao và khả năng ứng dụng rộng rãi hơn trên nhiều loại mạng khác nhau, như mạng xã hội, mạng học thuật hay mạng hợp tác doanh nghiệp.

TÀI LIỆU

[1] A. Dadlani, V. Vo, A. Khemka, S. Harvey, A. Kantoro, P. Jones, and D. Verhoeven, "Leading by the nodes: a survey of film industry network analysis and datasets," *Applied Network Science*, vol. 9, 12 2024.

[2] P. Wang, B. Xu, Y. Wu, and X. Zhou, "Link prediction in social networks: the state-of-the-art," *Science China Information Sciences*, vol. 58, 11 2014.

[3] S. Giri, S. Chaudhary, and B. Gautam, "Analyzing social networks of actors in movies and tv shows," 11 2024.

[4] M. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, pp. 167–256, 05 2003.

[5] D. Liben-nowell and J. Kleinberg, "The link prediction problem for social networks," *Journal of the American Society for Information Science and Technology*, vol. 58, 11 2003.

[6] M. Zhang and Y. Chen, "Link prediction based on graph neural networks," 02 2018.

[7] L. La Cava, D. Mandaglio, L. Zangari, and A. Tagarelli, "Heuristic-informed mixture of experts for link prediction in multilayer networks," 01 2025.

[8] J. A. Bondy, U. S. R. Murty *et al.*, *Graph theory with applications*. Macmillan London, 1976, vol. 290.

[9] M. E. J. Newman, "Scientific collaboration networks. i. network construction and fundamental results," *Phys. Rev. E*, vol. 64, p. 016131, Jun 2001.

[10] T. Zhou, L. Lü, and Y.-C. Zhang, "Predicting missing links via local information," *The European Physical Journal B - Condensed Matter and Complex Systems*, vol. 71, pp. 623–630, 10 2009.

[11] G. J. de Bruin, C. Veenman, H. Herik, and F. Takes, *Experimental Evaluation of Train and Test Split Strategies in Link Prediction*, 01 2021, pp. 79–91.

[12] I. Kalyani, A. Mathi, and N. Sett, "Evaluating link prediction: new perspectives and recommendations," *International Journal of Data Science and Analytics*, vol. 20, pp. 6855–6886, 07 2025.

[13] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," vol. 2016, 07 2016, pp. 855–864.

[14] T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 09 2016.

[15] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," vol. 06, 06 2006.