پیش‌بینی پیوند در شبکه‌های علمی با استفاده از یادگیری ماشین و گراف‌های وزن‌دار

سیدمهدی وحیدی پور[1*]، استادیار، علیرضا محمدی[1]

۱ دانشکده مهندسی برق و کامپیوتر، دانشگاه کاشان، کاشان، ایران، vahidipour@kashanu.ac.ir

چکیده: با سرعت گرفتن رشد علم و انتشار مقالات و افزایش زمینه‌های علمی، یافتن همکار پژوهشی مناسب، یافتن منابع تحقیق و زمینه تحقیق برای محققان و نهادهای مربوطه، روز به روز سخت‌تر می‌شود. با انتخاب درست این موارد، می‌توان بیشترین بازدهی را از هزینه و زمان صرف شده برای پژوهش کسب کرد. برای حل این مسئله می‌توان با ایجاد شبکه‌ای شامل مقالات، دانشمندان و سایر موجودیت‌های علمی و ارتباطات بین آن‌ها، یک شبکه علمی ایجاد کرد و با استفاده از پیش‌بینی پیوند ارتباطاتی که در آینده شکل می‌گیرد را پیش‌بینی کرد. در این مقاله چارچوبی مبتنی بر یادگیری ماشین برای پیش‌بینی پیوند در شبکه‌های علمی ارائه شده است. در این چارچوب با وزن‌دهی شبکه بر اساس زمان و محتوا، محاسبه ویژگی‌های ساختاری و متنی جاسازی شده و انتخاب و استخراج ویژگی انجام می‌شود. در نهایت نمونه‌گیری منفی با استفاده از خوشه‌بندی تولید می‌شود تا یک مدل یادگیری ماشین برای پیش‌بینی پیوند آموزش داده شود. هر یک از مراحل این چارچوب به صورت جدا و همه با هم آزمایش شدند و نتایج نشان داد روش وزن‌دهی پیشنهاد شده برای شبکه ارجاعات و همکاری نویسندگان باعث افزایش دقت معیارهای شباهت وزن‌دار و در نتیجه افزایش دقت کل الگوریتم می‌شود. همچنین نمونه‌گیری منفی با استفاده از خوشه‌بندی باعث بهتر آموزش داده شدن الگوریتم یادگیری ماشین می‌شود. ویژگی‌های متنی داده‌های علمی مانند عنوان و چکیده مقالات نیز نقش مؤثری در پیش‌بینی پیوندهای آینده دارند.

واژه‌های کلیدی: پیش‌بینی پیوند، شبکه ارجاعات، شبکه همکاری نویسندگان، یادگیری ماشین

[*] **سیدمهدی وحیدی پور، vahidipour@kashanu.ac.ir**

Link prediction in scientific networks using machine learning and weighted graphs

S. Mehdi Vahidipour[1*], Assistant Professor, Alireza Mohamadi

1 Electrical and Computer Engineering Faculty, University of Kashan, Kashan, Iran, vahidipour@kashanu.ac.ir.

Abstract: With the acceleration of the development of science and the publication of articles and the increase of scientific fields, finding suitable research partners, finding research sources and research fields for researchers and relevant institutions is becoming more and more difficult. By choosing these things correctly, you can get the most efficiency from the cost and time spent on research. To solve this problem, a scientific network can be created by creating a network including articles, scientists, and other scientific entities and the connections between them, and predicting the connections that will be formed in the future using link prediction. In this paper, a framework based on machine learning is presented for link prediction in scientific networks. In this framework, by weighting the network based on time and content, calculating embedded structural and textual features, feature selection and extraction, and finally negative sampling using clustering, a machine learning model is trained for link prediction. Each of the steps of this framework was tested separately and all together, and the results showed that the proposed weighting method for the network of references and authors' collaboration increases the accuracy of the weighted similarity criteria and, as a result, increases the accuracy of the entire algorithm. Also, negative sampling using clustering makes the machine learning algorithm better trained. The textual features of scientific data such as the title and abstract of articles also play an effective role in predicting future links.

Keywords: Link Prediction; Citation Networks; Author Collaboration Networks; Machine Learning.

*** S. Mehdi Vahidipour, vahidipour@kashanu.ac.ir**

۱- مقدمه

شبکه‌ها در زندگی امروزه ما نقش مهمی دارند و تحلیل شبکه‌ها اطلاعات مفیدی در اختیار ما قرار می‌دهد [1-2]. این اطلاعات این امکان را فراهم می‌کند که رفتار شبکه‌ها را بهتر درک کنیم و چگونگی تکامل آن‌ها را در آینده شبیه‌سازی کنیم. پیش‌بینی پیوند یکی از تحلیل‌هایی است که روی شبکه‌ها انجام می‌شود و امروزه اهمیت زیادی پیدا کرده است. پیش‌بینی پیوند به معنای یافتن پیوندهای از دست رفته و یا پیش‌بینی پیوندهای آینده است [۳-۴]. این پیش‌بینی می‌تواند بر اساس پیوندهای موجود در حال حاضر شبکه، ویژگی‌های رأس‌های شبکه و یا هر دو انجام شود [۵].

افراد جدیدی که یک شبکه اجتماعی پیشنهاد دنبال کردن آن‌ها را به کاربرانش می‌دهد مثالی از کاربرد پیش‌بینی پیوند در شبکه‌های اجتماعی است. اطلاعات گذشته کاربر در شبکه اجتماعی می‌تواند مبنای این پیشنهاد باشد؛ مثلاً افرادی را که دنبال می‌کند (ویژگی های ساختاری) و یا ویژگی‌هایی مانند شغل و محل سکونت (ویژگی‌های محتوایی) [۶].

پیش‌بینی پیوند تا کنون در زمینه‌های بسیاری مانند شبکه‌های اجتماعی، مسائل بیولوژیک و سیستم‌های توصیه‌گر به کار گرفته‌شده و منجر به نتایج قابل قبولی شده‌است [۵]. یکی از کاربردهای پیش‌بینی پیوند که به تازگی اهمیت پیدا کرده ولی کمتر مورد توجه قرار گرفته است، پیش‌بینی پیوند در شبکه‌های علمی است؛ شبکه‌های علمی نوع خاصی از شبکه‌های پیچیده[1] هستند که روابط بین موجودیت‌های علمی مانند دانشمندان، مقالات و نشریات را با مفاهیمی مانند ارجاع[2]، همکاری[3] و نویسندگی بازنمایی می‌کنند [۷].

با افزایش سرعت رشد علم و ایجاد زمینه‌های علمی جدید در سال‌های اخیر و همکاری دانشمندان در حوزه‌های مختلف علوم با یکدیگر، این نگرانی برای دانشمندان و سازمان‌های دانش‌بنیان به وجود آمده‌است که در سال‌های آینده، زمان و بودجه پژوهشی را به چه زمینه‌هایی اختصاص دهند. همچنین یافتن مقالات مرتبط با یک زمینه علمی خاص [۸] و یافتن دانشمندان متخصص برای همکاری‌های علمی [۹] از جمله مسائلی هستند که در شبکه‌های علمی و با استفاده از پیش‌بینی پیوند راه‌حل‌هایی برایشان ارائه شده است.

در پژوهش‌های انجام شده در زمینه شبکه‌های علمی، بیشتر تمرکز روی ساختار شبکه و چگونگی پیوند عناصر با یکدیگر بوده است و محتوای عناصر کمتر مورد توجه قرار گرفته است. در برخی پژوهش‌ها که محتوای متنی برای پیش‌بینی پیوند استفاده شده نیز از روش‌های ساده مانند شباهت کسینوسی بردار tf-idf استفاده شده است [۶]. همچنین در روش‌های یادگیری ماشین، با توجه به حجم زیاد شبکه‌های علمی، انتخاب داده‌های آموزشی به صورت تصادفی انجام شدند [۱۰]. در این مقاله برای بهبود این نقاط ضعف، روش‌های جاسازی کلمات[4] برای تحلیل محتوای متنی و استفاده از خوشه‌بندی برای انتخاب داده‌های آموزشی برای پیش بینی وزن پیشنهاد شده است. در نهایت با ترکیب روش‌های موجود و روش‌های پیشنهادی که نتایج را بهبود دادند، یک چارچوب کلی برای پیش‌بینی پیوند در شبکه‌های علمی پیشنهاد می‌کند.

[1] Complex Networks

[2] Citation

[3] Co-authorship

[4] Embedding

ساختار مقاله در ادامه بدین صورت است. در بخش دوم تعریف‌ها، اصطلاحات و مفاهیم پایه مربوط به شبکه‌های پیچیده، شبکه‌های علمی و پیش‌بینی پیوند شرح داده می‌شود و سپس چند الگوریتم و روش حل مسئله پیش‌بینی پیوند که مبنای روش‌های پیشنهادی و آزمایشات هستند بررسی خواهد شد. در بخش سوم کارهای پیشین در زمینه پیش‌بینی پیوند در شبکه‌های مبتنی بر مفاهیم علمی مورد بررسی قرار می‌گیرند. در بخش چهارم روش‌های پیشنهادی ارائه می‌شود و در فصل پنجم آزمایش‌هایی برای مقایسه روش‌های رایج موجود و روش‌های پیشنهادی طراحی می‌شوند. در بخش ششم، نتیجه‌گیری مقاله ارائه خواهد شد.

۲- مفاهیم پایه

ابتدا مفهوم شبکه‌های پیچیده بررسی شده و در ادامه، نوع خاصی از شبکه‌های پیچیده که به عنوان شبکه‌های علمی در این مقاله از آن یاد می‌شود، شرح داده می‌شود. سپس تعریف پیش‌بینی پیوند به طور کلی و کاربردهای آن در شبکه‌های علمی مرور می‌شوند. پس از آن رویکردها و روش‌های مورد استفاده در پیش‌بینی پیوند شرح داده شده است. در انتهای این بخش دو الگوریتم جاسازی کلمات برای تحلیل محتوای متنی و جاسازی رأس برای تحلیل ساختاری گراف تشریح شده است.

۲-۱ شبکه‌های پیچیده

امروزه شبکه‌ها به یکی از مدل‌های مهم برای بازنمایی[1] بسیاری از مسائل تبدیل شده‌اند. سیستم‌های اجتماعی، زیستی و اطلاعاتی بسیاری را می‌توان با شبکه‌ها توصیف کرد. رأس‌های این شبکه‌ها کاربران، کامپیوترها، افراد، عناصر زیستی (پروتئین، ژن و غیره) و مانند این‌ها هستند. یال‌ها نشانگر ارتباطات یا تعاملات بین رأس‌ها هستند [۵]. شبکه‌های دنیای واقعی که دارای ویژگی‌های ساختاری غیر بدیهی[2] هستند و تفاوت‌های زیادی با گراف‌های تصادفی و گراف‌های مشبک[3] دارند، شبکه‌های پیچیده هستند [۱۱]. شبکه هایی که پیش‌تر مثال زده‌شد، همگی شبکه‌های پیچیده هستند. با بررسی رفتار شبکه‌ها و پردازش آن‌ها، می‌توان به اطلاعات مفیدی دست یافت. تشخیص جامعه[4]، بصری‌سازی شبکه[5]، بررسی ساختار شبکه[6] و پیش‌بینی پیوند[7] از جمله پردازش‌هایی است که می‌تواند روی یک شبکه صورت گیرد [۴].

شبکه‌های پیچیده را می‌توان با داده‌ساختار گراف نشان داد. مدل کردن شبکه‌های پیچیده با استفاده از گراف این مزیت را فراهم می‌کند که از الگوریتم‌های پردازش گراف برای تحلیل و بررسی این شبکه‌ها استفاده کرد.

۲-۲ شبکه‌های علمی

در این مقاله، به شبکه‌هایی که متشکل از عناصر علمی مانند مقالات، نویسندگان، کلمات کلیدی، انتشارات و دانشگاه‌ها به همراه روابط بین آن‌ها هستند شبکه‌های علمی گفته می‌شود. شبکه های علمی چندان عبارت پر کاربردی در ادبیات پژوهش نیست و معمولاً در هر پژوهش از عناوین خاص منظوره برای نامیدن شبکه مورد بحث مانند شبکه همکاری نویسندگان[8] [۱۲] و شبکه‌ی ارجاعات (استنادات)[9] [۸, ۱۰] استفاده شده‌است. ولی با توجه به تعدد این شبکه‌ها که وجه مشترک همه‌ی آن‌ها «علم محور» بودن

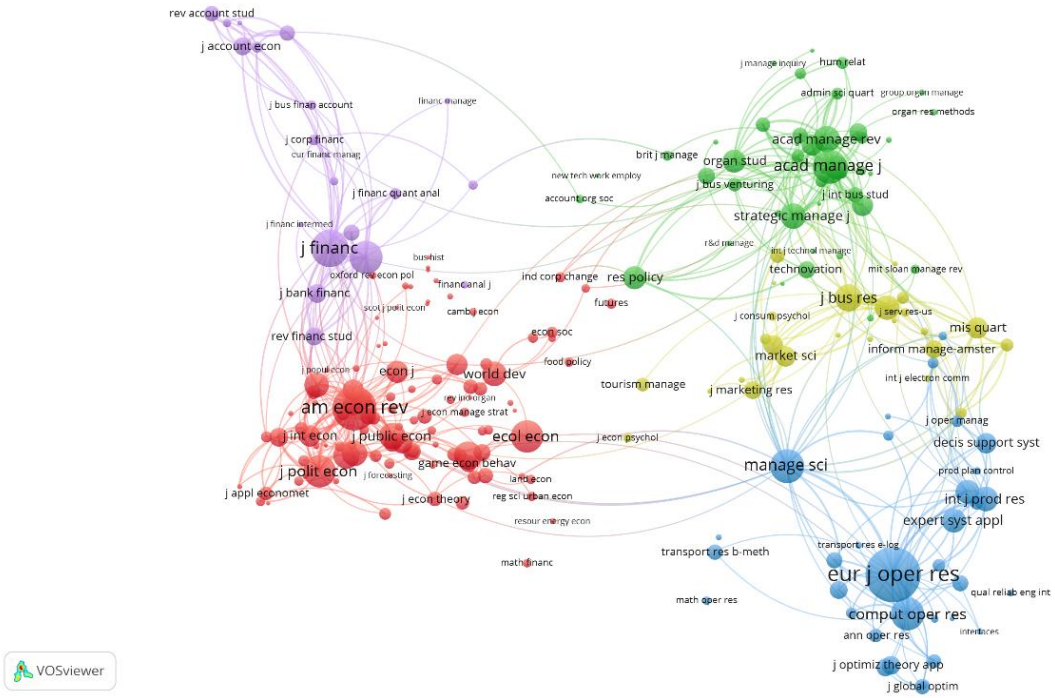
[1] representation

[2] Non-trivial
[3] lattice graph
[4] Community detection
[5] Network visualization
[6] Network structure analysis
[7] Link Prediction
[8] Co-authorship network
[9] Citation network

است از عبارت «شبکه‌های علمی» برای نامیدن این شبکه‌ها استفاده می‌کنیم.

در شکل ۱ یک شبکه علمی وجود دارد که متشکل از مجلات[1] علمی و روابط بین آن‌هاست. در صورتی که مقاله‌ای از یک مجله به مقاله دیگری از مجله دیگر استناد کرده باشد، بین این دو مجله یال برقرار خواهد شد؛ برای مشخص تر شدن شکل تمام اطلاعات نشان داده نشده است. همچنین رأس‌های همرنگ زمینه علمی یکسانی دارند و رأس‌های بزرگ‌تر مجلاتی هستند که تعداد مقاله بیشتری منتشر کرده‌اند.

[1] Journal



شکل ۱ نمونه‌ای از شبکه علمی ارتباط بین مجلات

شبکه‌های علمی می‌توانند وزن‌دارباشند. گراف‌های وزن دار گراف هایی هستند که در آن‌ها به هر یال عددی منتسب شده‌است که نمایانگر وزن آن یال است [۱۳]. مثلا در شبکه دانشمندان، دانشمندان رأس‌های شبکه همکاری بین دو دانشمند یال بین آنهاست. در چنین شبکه‌ای، وزن یال می‌تواند تعداد همکاری بین دو دانشمند را نشان دهد.

۲-۳ پیش‌بینی پیوند و کاربردهای آن در شبکه‌های علمی

اتصال بین دو رأس در یک شبکه پیچیده، پیوند نامیده می‌شود. یکی از بنیادی ترین مسائلی که روی شبکه‌های پیچیده مطرح می شود پیش‌بینی این پیوندهاست. در یک شبکه پیچیده با تعداد زیادی رأس، پیش‌بینی این که چه پیوندهایی در آینده تشکیل خواهند شد نقش مهمی در تحلیل و درک شبکه و نحوه تکامل آن دارد [۵]. اگر یک شبکه پیچیده را در زمان t فرض کنیم، که هر رأس نمایانگر یک موجودیت در شبکه و هر یال نشان‌دهنده تعامل رئوس باشد، پیش‌بینی یال‌هایی که در زمان $t + \Delta$ به شبکه اضافه خواهد شد مسئله پیش‌بینی پیوند نام دارد. به بیان ساده تر پیش‌بینی پیوندهایی که در آینده شکل خواهد گرفت یا در گذشته شکل گرفته بوده ولی از بین رفته[1] (و دوباره تشکیل خواهد شد)، پیش‌بینی پیوند نامیده می‌شود [۴]. در شبکه‌های علمی، پیش‌بینی پیوند می‌تواند به یافتن همکارهای پژوهشی مناسب [۱۲]، پیش بینی روند تولید دانش [۱۴] و یا پیش‌بینی معیارهای ارزیابی علمی مانند تعداد استنادات[2] [۹, ۱۵] کمک کند.

۲-۴ رویکردهای پیش‌بینی پیوند

[1] Missing links
[2] Citation count

از دو جنبه می‌توان گراف مورد مطالعه برای پیش‌بینی پیوند را بررسی کرد: ساختار گراف و محتوای گراف. رویکرد ساختاری گراف، فارغ از این که گراف چه پدیده‌ای از دنیای واقعی یا تئوری را بازنمایی می‌کند، صرفاً به ارتباط رأس‌های گراف می‌پردازد. در بسیاری از گراف‌های دنیای واقعی رأس‌ها و در بعضی موارد یال ها، داده‌هایی در خود دارند که نوع، ویژگی‌ها و مقادیر رأس یا یال را نشان می‌دهند. در رویکرد محتوایی گراف از این ویژگی‌ها برای پیش‌بینی پیوند استفاده می‌شود.

وزن یال را می‌توان جزئی از ساختار گراف دانست. همچنین می توان آن را یک ویژگی دانست که برای هر یال مقدار مشخصی دارد؛ در این صورت جزئی از محتوای گراف محسوب می‌شود. در این مقاله، فرض شده وزن جزئی از ساختار گراف است و معیارهای مشابهت وزن‌دار به عنوان راه حل پیش‌بینی پیوند در گراف‌های وزن‌دار بررسی شده است.

در ادامه هر یک از رویکردهای مبتنی بر ساختار و محتوا شرح داده شده‌است و در انتها روش‌هایی که رویکردهای ساختاری و محتوایی را ترکیب می‌کنند و از مزایای هر دو بهره می‌برند به طور خلاصه بیان شده است.

۲-۴-۱ رویکرد مبتنی بر ساختار گراف

در رویکردهای مبتنی بر ساختار گراف عموماً از معیارهای مشابهت برای پیش‌بینی پیوند استفاده می‌شود. معیار مشابهت بین دو رأس گراف محاسبه می‌شود و هر چقدر حاصل عددی بزرگتر باشد مشابهت بین دو رأس و در نتیجه احتمال ایجاد پیوند بالاتر است. معیارهای مشابهت به سه دسته محلی، شبه محلی و سراسری تقسیم می‌شوند.

معیارهای مشابهت محلی، تنها دو رأس هدف و دو لایه از همسایه های آن‌ها را در نظر می‌گیرند و بقیه‌ی قسمت‌های گراف تاثیری در نتیجه ندارند. به همین دلیل از روش‌های سراسری و شبه محلی پیچیدگی زمانی و مکانی کمتری دارند. با این وجود دقت قابل قبولی دارند و پژوهش‌ها نشان می‌دهد نسبت به معیارهای شبه محلی و سراسری کارآمدتر هستند [۴].

با توجه به حجم داده زیاد شبکه‌های علمی (بیش از صدهزار رأس و یک میلیون یال) مشابهت‌های شبه محلی و سراسری از نظر زمانی چندان کارامد و مقیاس پذیر نیستند. به همین دلیل، روش‌های پیشنهادی و آزمایشات مبتنی بر مشابهت در بخش‌های بعدی از مشابهت‌های محلی استفاده می‌کنند و منابع مطالعه شده نیز از این دسته مشابهت‌ها هستند.

در ادامه $S(x,y)$ به معنای مشابهت بین رأس x و y است، نماد Γ_x نمایانگر همسایه‌های رأس x می‌باشد و $w(x,y)$ به معنای وزن یالی است که بین رئوس x و y قرار دارد.

همسایه مشترک[1]

ساده‌ترین معیار مشابهت در پیش‌بینی پیوند همسایه مشترک است، همسایه مشترک به صورت تعداد همسایه‌های مشترک دو رأس تعریف می‌شود. این معیار با وجود سادگی، در کاربردهای دنیای واقعی نتیجه مطلوبی کسب می‌کند و پایه معیارهای پیچیده تر است [۴]. در این معیار فرض می‌شود هر چقدر دو رأس همسایه های مشترک بیشتری داشته‌باشند، احتمال این که خود نیز با یکدیگر همسایه شوند بیشتر است (معادله ۱) [۱۶]. در نسخه وزن دار که در معادله ۲ نشان داده شده است، به جای در نظر گرفتن

[1] Common neighbors

تعداد همسایه‌های مشترک، مجموع وزن همسایه‌های مشترک تا دو رأس x و y محاسبه می‌شود [۱۷].

$$s_{CN}(x,y) = |\Gamma_x \cap \Gamma_y| \quad ۱$$

$$s_{WCN}(x,y) = \sum_{z \in \Gamma_x \cap \Gamma_y} w(x,z) + w(z,y) \quad ۲$$

آدامیک آدار[1]

این معیار، بهبود یافته معیار همسایه مشترک است، در معیار آدامیک آدار (معادله ۳)، هر همسایه مشترک با لگاریتم درجه‌اش جریمه می‌شود. یعنی همسایه‌های مشترکی که با رأس‌های زیادی یال برقرار کردند ارزش کمتری نسبت به همسایه‌های مشترکی دارند که با رأس‌های کمتری پیوند دارند [۱۸]. در نسخه وزن‌دار آدامیک آدار کافیست به جای جریمه‌کردن با لگاریتم درجه همسایه مشترک، لگاریتم مجموع وزن‌های یال‌های هر همسایه مشترک را جریمه کنیم [۶] (به معادله ۴ رجوع کنید).

$$s_{AA}(x,y) = \sum_{z \in \Gamma_x \cap \Gamma_y} \frac{1}{log|\Gamma_z|} \quad ۳$$

$$s_{WAA}(x,y) = \sum_{z \in \Gamma_x \cap \Gamma_y} \frac{w(x,z) + w(z,y)}{log \sum_{z' \epsilon \Gamma_z} w(z',z)} \quad ۴$$

ژاکارد

معیار ژاکارد یک معیار قدیمی و بسیار پرکاربرد در زمینه‌های مختلف است که اولین بار برای بررسی مشابهت مجموعه‌ها مورد استفاده قرار گرفت. این معیار نسبت همسایه‌های مشترک به تمام همسایه‌های دو رأس را حساب می‌کند. این معیار نیز بهبود یافته همسایه مشترک است که تعداد همسایه‌های مشترک با استفاده از تعداد کل همسایه‌ها جریمه می‌شوند. در واقع اگر دو جفت رأس تعداد همسایه‌های مشترکی داشته باشند، جفت رأسی که تعداد کل همسایه‌های کمتری دارد احتمال بیشتری برای برقراری پیوند دارد. تعریف استاندارد معیار ژاکارد در معادله ۵ و تعریف وزن دار آن در معادله ۶ آمده است [۱۷].

$$s_{JA}(x,y) = \frac{|\Gamma_x \cap \Gamma_y|}{|\Gamma_x \cup \Gamma_y|} \quad ۵$$

$$s_{JA}(x,y) = \frac{\sum_{z \in \Gamma_x \cap \Gamma_y} w(x,z) + w(z,y)}{\sum_{a \in \Gamma_x} w(a,x) + \sum_{b \in \Gamma_y} w(b,y)} \quad ۶$$

تخصیص منابع

این شاخص از فرایند تخصیص منابع که در شبکه‌های پیچیده صورت می‌گیرد الهام گرفته شده است. شاخص تخصیص منابع، انتقال واحدهای منابع بین دو رأس غیر متصل x و y از طریق همسایگان آن‌ها را مدل می‌کند، هر رأس همسایه یک واحد از منابع را از رأس x دریافت می‌کند و آن را به صورت مساوی بین همسایگانش توزیع می‌کند. میزان منابع دریافت شده توسط y به عنوان مشابهت بین این دو رأس در نظر گرفته می‌شود [۱۹]. در نسخه وزن‌دار این شاخص، توزیع منابع از رأس x به همسایگان به نسبت وزن یال هر همسایه صورت می‌گیرد [۱۷].

$$s_{RA}(x,y) = \sum_{z \in \Gamma_x \cap \Gamma_y} \frac{1}{|\Gamma_z|} \quad ۷$$

$$s_{WRA}(x,y) = \sum_{z \in \Gamma_x \cap \Gamma_y} \frac{w(x,z) + w(z,y)}{\sum_{z' \epsilon \Gamma_z} w(z',z)} \quad ۸$$

اتصال ترجیحی

شاخص اتصال ترجیحی بر این مبنا کار می‌کند که رأس‌هایی که پیوندهای زیادی تشکیل داده‌اند با احتمال بیشتری با یکدیگر پیوند تشکیل می‌دهند. مقدار این شاخص برای دو رأس برابر با حاصلضرب تعداد همسایه‌های دو رأس است [۲۰]. این شاخص بر خلاف شاخص‌های پیشین بر مبنای همسایه‌های مشترک کار

[1] Adamic-Adar

نمی‌کند و معمولا در پژوهش‌ها نتایج دقیقی از آن حاصل نمی‌شود ولی با توجه به پیچیدگی زمانی کمتر از شاخص‌های مبتنی بر همسایه مشترک، می‌تواند در کنار آن به عنوان مکمل استفاده شود [۴]. نسخه وزن‌دار این شاخص، حاصلضرب مجموع یال‌های دو رأس‌را به عنوان مقدار شاخص محاسبه می‌کند [۱۷].

$$S_{PA}(x,y) = |\Gamma_x| \times |\Gamma_y| \qquad ۹$$

$$S_{WPA}(x,y) = \sum_{x' \in \Gamma_x} (x,x') \times \sum_{y' \in \Gamma_y} (y,y') \qquad ۱۰$$

۲-۴-۲ رویکرد مبتنی بر محتوا و زمان

معیارهای مشابهت گفته شده صرفاً ساختار شبکه را در نظر می گیرند. منظور، ساختار[1] دو رأسی است که شباهتشان حساب می شود و همسایه‌های آن‌هاست. در صورتی که علاوه بر ساختار گراف، موارد دیگری نیز وجود دارند که می‌توان از آن‌ها برای پیش بینی پیوند کمک گرفت. به عنوان مثال در یک شبکه اجتماعی ممکن است دو نفر هیچ همسایه مشترک و در نتیجه مشابهت ساختاری نداشته باشند ولی به واسطه شغلشان تمایل به برقراری پیوند داشته باشند و در آینده پیوند برقرار کنند، در اینجا شغل، محتوای رئوس در نظر گرفته می‌شود و مستقل از ساختار است.

همچنین زمان برقراری پیوند در گراف می‌تواند حائز اهمیت باشد. مثلاً در شبکه همکاری نویسندگان می‌توانیم این فرضیه را مطرح کنیم که هر چقدر زمان به وجود آمدن یک پیوند بیشتر باشد (سابقه همکاری بیشتر باشد)، آن پیوند اهمیت بیشتری دارد. به طورکلی، با ترکیب اطلاعات زمانی و محتوایی با ساختار شبکه، نتیجه پیش‌بینی پیوند بهبود چشمگیری می‌یابد [۶].

۲-۴-۳ رویکردهای ترکیبی

ساختار و محتوای گراف هر دو داده‌های ارزشمندی برای پیش‌بینی پیوند محسوب می‌شوند. روش‌هایی مانند مشابهت‌ها که فقط ساختار گراف را در نظر می‌گیرند یا روش‌های مبتنی بر محتوا که فقط بر محتوا تمرکز دارند و ساختار را نادیده می‌گیرند در مسائلی که هم محتوا و هم ساختار اهمیت زیادی دارند منجر به نتیجه قابل قبولی نمی‌شوند. به همین خاطر روش‌هایی ارائه شده‌اند که ساختار و محتوا را هم‌زمان در پیش بینی پیوند لحاظ می‌کنند.

دسته‌ای از این روش‌ها که به عنوان روش‌های مبتنی بر وزن‌دار کردن شناخته می‌شوند، محتوای گراف را در قالب وزن در ساختار گراف تعبیه می‌کنند و سپس با روش‌های مبتنی بر ساختار مانند معیارهای مشابهت وزن‌دار به حل مسئله پیش‌بینی پیوند در گراف های دارای محتوا می‌پردازند [۶, ۱۲].

دسته‌ای از روش‌های دیگر نیز برای هر یال بردارهایی ایجاد می کنند که درایه‌های آن مقادیر ویژگی‌های ساختاری و محتوایی گراف است. این بردارها را با استفاده از معیارهای مشابهت برداری یا الگوریتم‌های یادگیری ماشین پردازش می‌کنند و خروجی آن را برای حل مسئله پیش‌بینی پیوند استفاده می‌کنند [۱۰, ۱۴, ۲۱].

۲-۵ جاسازی کلمات[2]

در پردازش زبان طبیعی[3] دسته‌ای از روش‌ها وجود دارند که هر کلمه را به یک بردار با طول ثابت نگاشت می‌کنند. الگوریتم‌های زیادی هستند که وظیفه نگاشت هر کلمه به یک بردار را انجام می دهند که از نظر روش انجام کار با یکدیگر تفاوت دارند ولی وجه مشترک همه این‌الگوریتم‌ها تبدیل کلمه به یک بردار به طول ثابت

[1] Topology, Structure

[2] Word embedding

[3] Natural Language Processing (NLP)

است. این روش‌ها از این جهت سودمند هستند که به راحتی می توان عملیات محاسباتی[1] را روی کلمات، جملات و پرونده‌ها انجام داد. با تبدیل متن به بردار، این امکان فراهم می‌شود که عملیاتی مانند جمع، تفریق، محاسبه فاصله و محاسبه مشابهت را از طریق عملیات محاسباتی برداری انجام دهیم [۲۲].

الگوریتم‌های word2vec و GloVe[2] از جمله الگوریتم‌های مطرح در این زمینه هستند که منجر به نتایج خوبی شده‌اند [۲۳]. از آن‌جایی که بخش بزرگی از محتوای شبکه‌های علمی مانند عنوان مقاله و چکیده در قالب متن می‌باشد، در روش‌های پیشنهادی از الگوریتم‌های جاسازی کلمات برای پردازش محتوای متنی شبکه استفاده خواهد شد.

۲-۶ جاسازی رأس‌ها[3]

پردازش ساختار داده گراف با استفاده از الگوریتم‌های استاندارد یادگیری ماشین و داده‌کاوی کار مشکلی است، به این خاطر که ورودی این الگوریتم‌ها داده‌های عددی ساختار یافته هستند ولی گراف، متشکل از تعدادی رأس و ارتباطات میان این رأس‌هاست [۲۴, ۲۵]. برای ایجاد نگاشت بین داده‌های رابطه‌ای مانند گراف و داده‌های ساختار یافته که قابلیت انجام عملیات محاسباتی روی آن ها فراهم باشد الگوریتم‌هایی ارائه شده‌است. یکی از این الگوریتم ها که در بسیاری از مسائل گراف مانند پیش‌بینی پیوند، تشخیص جامعه و دسته‌بندی رئوس نتایج خوبی به دست آورده‌است الگوریتم node2vec است [۲۶]. برای مطالعه جزییات این روش می‌توانید به پیوست الف مراجعه کنید.

۳- مرور کارهای گذشته

پیش‌از این پژوهش‌هایی در زمینه پیش‌بینی پیوند در شبکه‌های علمی انجام شده که در هر یک از این پژوهش‌ها یک مسئله روی یک نوع شبکه خاص (داده مورد مطالعه) برای یک کاربرد مشخص تعریف شده و با یک روش پیشنهادی پاسخی به آن مسئله داده شده است. بنابراین، مرور کارهای گذشته از دو جنبه کاربرد و شبکه مورد مطالعه و بررسی می‌شوند.

۳-۱ شبکه‌ها و کاربردهای مورد مطالعه

یکی از وجوه تمایز مهم پژوهش‌های انجام شده در زمینه پیش‌بینی پیوند شبکه‌های علمی، شبکه‌ای (گرافی) است که مورد پردازش قرار گرفته است. با مقایسه این پژوهش‌ها یک دسته‌بندی کلی از انواع شبکه‌های مورد پردازش و کاربردهایی از دنیای واقعی که برای پیش‌بینی پیوند روی آن شبکه تعریف شده ارائه می‌شود.

---

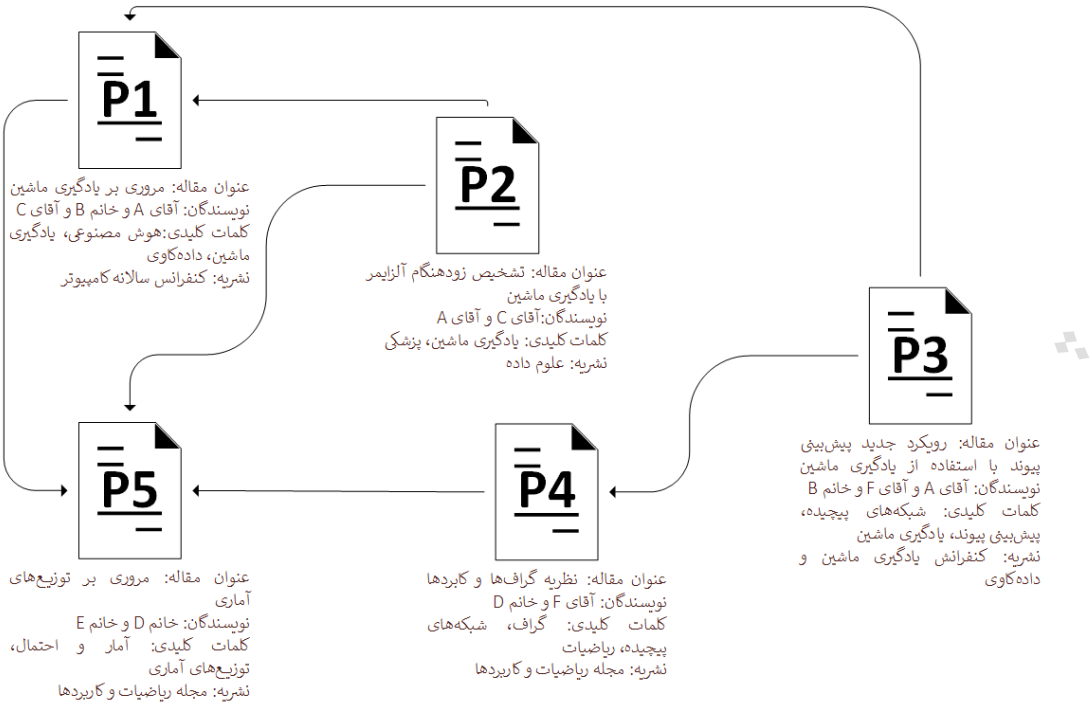
[1] Arithmetic
[2] Global Vectors for word representation
[3] Node embedding

شکل ۲ شبکه‌های پر کاربرد در پیش‌بینی پیوند شبکه‌های علمی

جدول ۱: مجموعه داده‌های مورد استفاده در کارهای پیشین

نام مجموعه داده	تعداد رأس‌ها	تعداد پیوندها	نوع شبکه
arXiv API[1]	به روز میشود		شبکه ارجاعات
DBLP[2]	به روز میشود		شبکه ارجاعات
HEP-TH[3]	27،770	352،807	شبکه ارجاعات
SCI- Expanded (Web of Science)[4]	۵۳ میلیون	۱ میلیارد	شبکه ارجاعات
Citeseer[5]	3،312	4،732	شبکه ارجاعات
ogbl-citation2[6]	2,927,963	30,561,187	شبکه ارجاعات
ogbl-collab	235,868	1,285,465	شبکه همکاری نویسندگان

[1] https://info.arxiv.org/help/api/
[2] https://dblp.org/
[3] https://snap.stanford.edu/data/cit-HepTh.html
[4] https://clarivate.com/webofsciencegroup/solutions/webofscience-scie/
[5] https://relational.fit.cvut.cz/dataset/CiteSeer
[6] https://ogb.stanford.edu/docs/linkprop/



شکل ۳ نمونه‌ای از شبکه ارجاعات

جامع ترین شبکه علمی موجود در دسترس، شبکه ارجاعات است که در صورت کامل بودن فرا داده‌های آن (نام نویسندگان، کلمات کلیدی و غیره) می‌توان برخی شبکه‌های علمی دیگر مانند شبکه کلمات کلیدی و شبکه همکاری نویسندگان را بر اساس آن ایجاد کرد (شکل ۲). مواردی که در شکل با رنگ آبی مشخص شدند در مقالات مطالعه شده صریحاً مورد استفاده قرار گرفته‌اند و مواردی که با رنگ خاکستری نشان داده شده‌اند، مواردی هستند که در پژوهش‌های پیشین مورد مطالعه ما به آن‌ها اشاره نشده ولی از طریق مجموعه‌داده‌های موجود قابل به دست آمدن هستند و می توانند بالقوه مفید باشند. در بخش‌های بعدی هر یک از شبکه‌هایی که در پژوهش‌های پیشین از آن استفاده شده بررسی می‌شود.

با توجه به ویژگیهای مشترک شبکه‌ها، این دسته‌بندی برای یافتن الگوریتم مناسب پیش‌بینی پیوند مفید است . علاوه بر این، به ایجاد دید جامع نسبت به شبکه‌هایی که در این مقاله مورد مطالعه قرار خواهند گرفت کمک خواهد کرد.

با توجه به ویژگیهای مشترک شبکه‌ها، این دسته‌بندی برای یافتن الگوریتم مناسب پیش‌بینی پیوند مفید است. علاوه بر این، به ایجاد دید جامع نسبت به شبکه‌هایی که در این مقاله مورد استفاده خواهند بود کمک خواهد کرد. در جدول ۱ مجموعه شبکه علمی در دنیای واقعی که در کارهای پیشین مورد استفاده قرار گرفتند به همراه تعداد رأس‌ها و پیوندها نام برده شده است. مجموعه داده ogbl-collab شبکه همکاری نویسندگان و بقیه شبکه ارجاعات هستند.

۳-۱-۱ شبکه ارجاعات

شبکه ارجاعات، جامع ترین شبکه علمی است که در آن مقاله‌ها رأس‌ها و ارجاعات یال‌های شبکه هستند. اگر مقاله‌ای به مقاله دیگری ارجاع داده باشد، میان دو رأس متناظر یال برقرار می‌شود.

مجموعه‌داده‌هایی مانند DBLP در دسته شبکه ارجاعات قرار می گیرند. در شکل ۳ نمونه‌ای از شبکه ارجاعات مشاهده می‌شود. در این شبکه، مقاله P2 به مقاله P1 ارجاع داده است، بنابراین یک یال جهت دار از مقاله P2 به P1 در شبکه تشکیل می‌شود. به همین ترتیب سایر یال‌های شبکه بر اساس ارجاعات بین مقالات ساخته می‌شوند.

در مقاله [۱۲]، پیش‌بینی ارجاع متقابل با استفاده از یادگیری ماشین روی ویژگی‌های ساختاری و متنی شبکه ارجاعات Citeseer انجام شده است. پیش‌بینی تعداد ارجاعات به یک مقاله نیز بر اساس ویژگی‌های ساختاری در مجموعه داده HEP-TH با استفاده از یادگیری ماشین در مقاله [۶] انجام شده است. در مقاله [۵] روی شبکه ارجاعات TEP-TH، برای مسئله یافتن مقالات مرتبط با یک مقاله، یک راه حل با استفاده از معیارهای شباهت محلی ارائه شده است. مقاله [۷] نیز با استفاده از tf-idf چکیده مقالات و ساختار شبکه، روی پنج پایگاه داده ارجاعات SCI- Expanded پیش‌بینی پیوند انجام داده است.

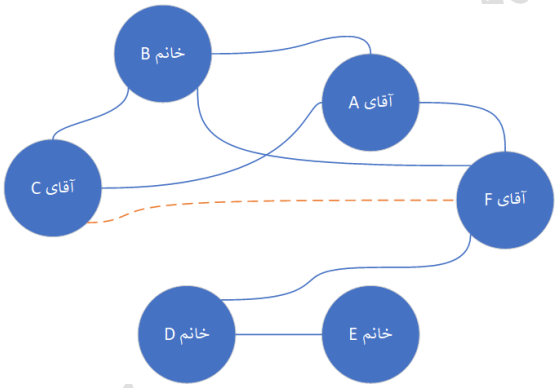
۳-۱-۲ شبکه همکاری نویسندگان

در این شبکه، نویسندگان رأس‌های شبکه را تشکیل می‌دهند و در صورت انجام همکاری بین دو نویسنده، بین آن‌ها یال تشکیل می شود. شبکه همکاری نویسندگان می‌تواند وزن‌دار باشد؛ وزن هر یال نشان دهنده تعداد همکاری دو نویسنده است.

با توجه به شبکه ارجاعات شکل ۳، «آقای A»، «خانم B» و «آقای C» نویسنده‌های همکار در مقاله P1 هستند. به همین خاطر بین آن ها یال برقرار می‌شود، به همین ترتیب برای سایر مقالات، نویسنده ها به گراف اضافه می‌شوند و بین نویسنده‌های مشترک یک مقاله یال برقرار می‌شود. در نهایت شبکه نویسندگان به صورت شکل ۴ ساخته می‌شود.

در مقاله‌های [۳, ۹] پیش‌بینی همکاری نویسندگان با استفاده از یک روش ترکیبی مبتنی بر وزن‌دار کردن شبکه همکاری با مشابهت محتوایی و سپس پیش‌بینی پیوند با معیارهای شباهت وزن‌دار با استفاده از شبکه‌های همکاری نویسندگان استخراج شده از HEP-TH و arXive انجام شده است. در مقاله [۲۷] نیز با افزودن کلمات کلیدی به شبکه همکاری نویسندگان مستخرج از شبکه SCI-Expanded و سپس اعمال الگوریتم قدم‌زدن تصادفی، پیش‌بینی انجام شده است.

پیش‌بینی پیوند در شبکه همکاری نویسندگان، افرادی را به هم معرفی می‌کند که می‌توانند در پژوهش‌های آتی همکاری مؤثر داشته باشند. برای مثال پیش‌بینی همکاری آتی بین «آقای F» و «آقای C» در شکل ۴ با نقطه چین قرمز مشخص شده است.



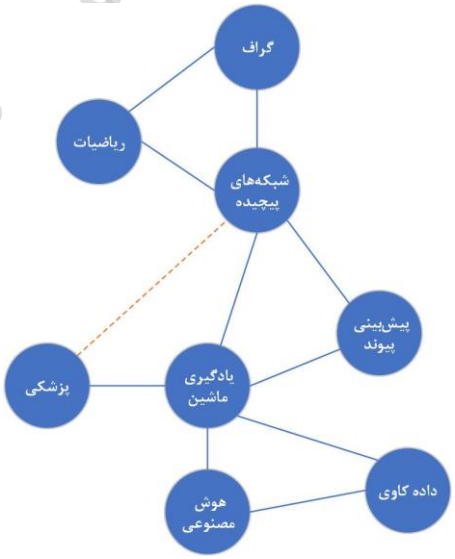
شکل ۴ شبکه نویسندگان مبتنی بر شکل ۳

۳-۱-۳ شبکه کلمات کلیدی

رأس‌های این شبکه، کلمات کلیدی مقالات هستند و در صورتی که دو کلمه کلیدی در یک مقاله استفاده شده باشند، بین آن‌ها یال ایجاد می‌شود. همان‌طور که پیش‌تر اشاره‌شد در صورت وجود

فراداده‌های مربوط به کلمات‌کلیدی مقاله در شبکه ارجاعات، می توان شبکه کلمات‌کلیدی از روی شبکه ارجاعات ایجاد کرد.

در مقاله [۱۴] برای پیش‌بینی روند تولید علم، با یک روش یادگیری ماشین و انتخاب ویژگی، روی شبکه کلمات کلیدی پیش‌بینی پیوند انجام شده است. به عنوان مثال در شکل ۵ در صورتی که الگوریتم پیش‌بینی پیوند، ایجاد پیوند بین دو کلمه کلیدی «شبکه‌های پیچیده» و «پزشکی» را پیش‌بینی کند می‌توان انتظار داشت در آینده مقالاتی که در زمینه ترکیب شبکه‌های پیچیده و پزشکی منتشر شوند مورد توجه قرار خواهند گرفت.



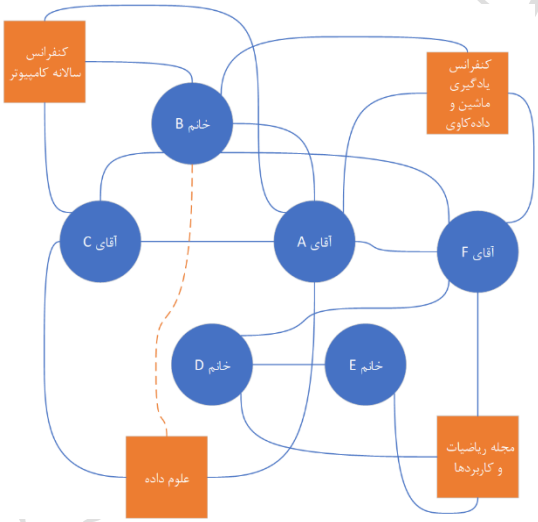
شکل ۵ شبکه‌ی کلمات کلیدی

۳-۱-۴ سایر شبکه‌های علمی

با توجه به داده‌هایی که در شبکه ارجاعات وجود دارد، می‌توان شبکه‌های دیگری ایجاد کرد که در منابع مطالعه شده به آن‌ها پرداخته نشده. شبکه نشریات و شبکه دانشگاه‌ها مثال‌هایی هستند که در شکل ۲ با رنگ خاکستری مشخص شده‌اند. ایجاد این شبکه ها و انجام پیش‌بینی پیوند در آن‌ها می‌تواند ارزشمند باشد.

دسته دیگری از شبکه‌های علمی شبکه‌های ناهمگن[1] که رأس‌های آن موجودیت‌هایی از انواع مختلف هستند. برای مثال موجودیت نشریات، نویسندگان، کلیمات کلیدی و ارتباط بین آن‌ها نمونه‌ای از شبکه علمی ناهمگن است (شکل ۶). در این شبکه دو موجودیت نویسنده و نشریه وجود دارد که نویسندگان با یکدیگر و نشریات ارتباط دارند.

پیش‌بینی پیوند در شبکه‌های ناهمگن می‌تواند بین رأس‌های هم نوع یا متفاوت صورت گیرد. در مقاله [۲۷] با افزودن کلمات کلیدی به شبکه همکاری نویسندگان، یک شبکه ناهمگن ایجاد می‌شود که باعث بهبود نتیجه پیش‌بینی پیوند بین نویسندگان شده است. در شکل ۶ نقطه چین قرمز نشان دهنده یک پیش‌بینی پیوند بین نویسنده و نشریه است؛ نویسنده مورد نظر در صورت چاپ مقاله خود در نشریه پیش‌بینی شده می‌تواند موفقیت بیشتری کسب کند.



شکل ۶ یک گراف ناهمگن متشکل از نویسنده‌ها و نشریات و ارتباط بینشان

۳-۲ روش‌های حل مسئله پیش‌بینی پیوند

[1] non-homogeneous

در این بخش مروری بر روش‌های پرکاربرد پیش‌بینی پیوند در شبکه‌های علمی خواهد شد و مزایا و معایب هر روش بیان می شود. در شکل ۷ یک دسته‌بندی از راه‌حل‌های پیش‌بینی پیوند مورد استفاده در مقالات مروری و منابع مربوط به پیش‌بینی پیوند در شبکه‌های علمی مورد مطالعه آورده شده‌است. در این دسته بندی موارد آبی رنگ صریحاً در منابع مطالعه شده در زمینه شبکه‌های علمی استفاده شده‌اند، موارد سفید رنگ در مقالات مروری مربوط به پیش‌بینی پیوند به کار رفته‌اند [۴, ۵] و برای کامل بودن نمودار آورده شده‌اند.

شکل ۷ راه‌حل‌های مسئله پیش‌بینی پیوند در شبکه‌های علمی

۳-۲-۱ مبتنی بر مشابهت

روش‌های مبتنی بر مشابهت، یکی از رایج‌ترین روش‌های پیش‌بینی پیوند هستند. جفت رأس‌هایی که امکان وجود پیوند برای آن‌ها قرار است بررسی شود، از بین تمام جفت رأس‌های موجود که پیوندی بین آن‌ها وجود ندارد مقدار معیار مشابهت حساب می شود. جفت رأسی که مقدار مشابهت‌شان بیشتر از حد آستانه باشد به عنوان پیوند پیش‌بینی شده در نظر گرفته می‌شوند.

در مقاله [۸] مقدار آستانه به صورت پویا محاسبه شده است، به این ترتیب که در شبکه اولیه مشابهت تمام جفت رئوسی که پیوند بین آن‌ها برقرار است محاسبه می‌شود و میانگین این مشابهت به عنوان حد آستانه در نظر گرفته می‌شود. در پژوهش [۴] مشابهت‌ها بر اساس زیاد به کم مرتب شده و K یال که بالاترین مشابهت را دارد به عنوان پیوند پیش‌بینی شده انتخاب شده‌اند. همچنین در مقالات [۱۰, ۱۴] از معیارهای مشابهت محلی به همراه داده‌های ساختاری و محتوایی دیگر برای ورودی الگوریتم یادگیری ماشین استفاده شده ولی مستقیماً مورد استفاده قرار نگرفته است.

روش مشابهت محلی ساده‌ترین روش پیش‌بینی پیوند است. از جمله معایب این روش می‌توان به عدم امکان استفاده همزمان از چند معیار مشابهت اشاره کرد. از طرف دیگر معیارهای مشابهت صرفاً روی ساختار گراف تعریف شده‌اند و در صورت نیاز به ترکیب زمان و محتوا با ساختار، این روش غیر قابل استفاده است. این معایب تا حدودی در روش وزن‌دار کردن و به طور کامل در روش مبتنی بر یادگیری ماشین رفع شده‌است. در ادامه کارهای پیشین مبتنی بر روش‌های وزن‌دار کردن و یادگیری ماشین بررسی شده‌اند.

۳-۲-۲ مبتنی بر وزن‌دار کردن

شبکه‌های وزن‌دار در همه کاربردهای دنیای واقعی از جمله شبکه های اجتماعی می‌تواند وجود داشته باشد و منحصر به شبکه‌های علمی نیست ولی از آن‌جایی که برخی از پژوهش‌ها در زمینه شبکه های علمی منحصراً روی شبکه‌های وزن‌دار پژوهش است، در این بخش جداگانه به آن‌ها پرداخته می‌شود. در پژوهش‌های پیشین مربوط به شبکه‌های علمی وزن‌دار، برخی از شبکه‌ها ذاتاً وزن دار بوده‌اند و در برخی دیگر با استفاده از روشی یک وزن به یال‌ها نگاشت شده. است. مقاله [۲۸] نشان می‌دهد با حذف وزن یال‌ها از گراف‌های وزن‌دار، نتیجه به صورت چشم‌گیری تغییر می‌کند و دقت کاهش می‌یابد.

در روش‌های مبتنی بر وزن‌دار کردن، با در نظر گرفتن ساختار، محتوا و برخی موارد نیز زمان، برای هر رأس یک وزن محاسبه می کنند و سپس با استفاده از معیارهای مشابهت وزن‌دار مانند آن‌چه پیش‌تر گفته شد، وجود یا عدم وجود پیوند بین رئوس را پیش‌بینی می‌کنند. در این روش، مشابهت ساختاری، مشابهت محتوایی و زمان در یک عدد که وزن یال است خلاصه می‌شود.

یک نمونه‌از رابطه‌های وزن‌دار کردن گراف در معادله ۱۱ قابل مشاهده است، توسط مقاله [۶] روی شبکه همکاری نویسندگان ارائه شده‌است، این رابطه از ضرب سه مقدار در یکدیگر محاسبه می‌شود. α و β دو ضریب هستند به طوری که $\alpha + \beta = 1$؛ با افزایش (کاهش) تأثیر محتوا، تأثیر زمان کم (زیاد) می‌شود. $|E(u,v)|$ وزنی‌است که یال گراف از ابتدا داشته، در گراف‌های بدون وزن این مقدار ۱ است. در قسمت دوم $\beta^{\frac{CTime-\max(t_{(u,v)})}{CTime-\min(t)}}$، $CTime$ زمان فعلی، $min(t)$ کمترین زمان دیده شده در شبکه (زمان به وجود آمدن شبکه) و $\max(t_{(u,v)})$ از بین زمان به وجود آمدن رأس u و v ماکسیمم را انتخاب می‌کند. دلیل این که $CTime - \max(t_{(u,v)})$ بر $CTime - \min(t)$ تقسیم شده، این است که مقدار کل نرمال شده و عددی بین ۰ و ۱ شود. به این ترتیب با ضرب این سه مقدار عددی به دست می‌آید که از ساختار، زمان و محتوا تاثیر گرفته است و به عنوان وزن یال u, v در نظر گرفته می‌شود. در ادامه با استفاده‌از معیارهای مشابهت وزن‌دار، مشابهت بین u, v محاسبه می‌شود و اگر بالاتر از حد آستانه باشد، بین آن‌ها پیوند برقرار می‌شود. نتایج پژوهش نشان می‌دهد با این روش، نتایج نسبت به هنگامی که فقط ساختار یا فقط محتوا در نظر گرفته شده بهبود می‌یابد.

$$w^{CTT}(u,v) = |E(u,v)| * \beta^{\frac{CTime-max(t_{(u,v)})}{CTime-min(t)}} * \alpha^{1-cos(u,v)} \quad ۱۱$$

در مقاله [۱۲] نیز با یک معیار مشابهت جدید به نام LDAcosine مشابهت کسینوسی دو رأس با در نظر گرفتن ساختار و محتوا محاسبه می‌شود و نتایج بهتری نسبت به روش‌های بدون محتوا به دست آمده است.

۳-۲-۳ مبتنی بر یادگیری ماشین

در پژوهش‌هایی که مسئله پیش‌بینی پیوند را با الگوریتم‌های یادگیری ماشین حل کردند یک رویکرد کلی مشاهده می‌شود. در همه این پژوهش‌ها برای هر یال تعدادی ویژگی محاسبه شده و هر یال به یک بردار چند بعدی تبدیل شده، هر یال با توجه به این که در گراف آموزشی وجود دارد یا خیر برچسب وجود یا عدم وجود پیوند می‌گیرد و به این ترتیب مسئله پیش‌بینی پیوند به یک مسئله دسته‌بندی تبدیل می‌شود.

در مقاله [۱۰] روی شبکه ارجاعات، برای یال‌های ممکن یک بردار متشکل از ۱۱ ویژگی محاسبه شده، برخی از این ویژگی‌ها مانند

همسایه مشترک و ژاکارد معیارهای ساختاری و برخی مانند اختلاف سال انتشار ۲ مقاله و تعداد نویسندگان مشترک ویژگی هایی محتوایی بودند. در انتها بردارها به الگوریتم ماشین بردار پشتیبان داده شده و نتیجه نشان دهنده بهبود نسبت به روش‌های پیشین روی مجموعه‌داده‌های بررسی شده است.

همچنین در پژوهشی که روی شبکه کلمات کلیدی انجام گرفته [۱۴] و از میان ۱۲ ویژگی محاسبه شده برای یال‌ها، ویژگی‌های ژاکارد، تخصیص منابع[1]، میانگین ضریب خوشگی[2] و میانگین مرکزیت[3] دو رأس انتخاب شدند؛ بهترین نتیجه با الگوریتم جنگل تصادفی به دست آمده است.

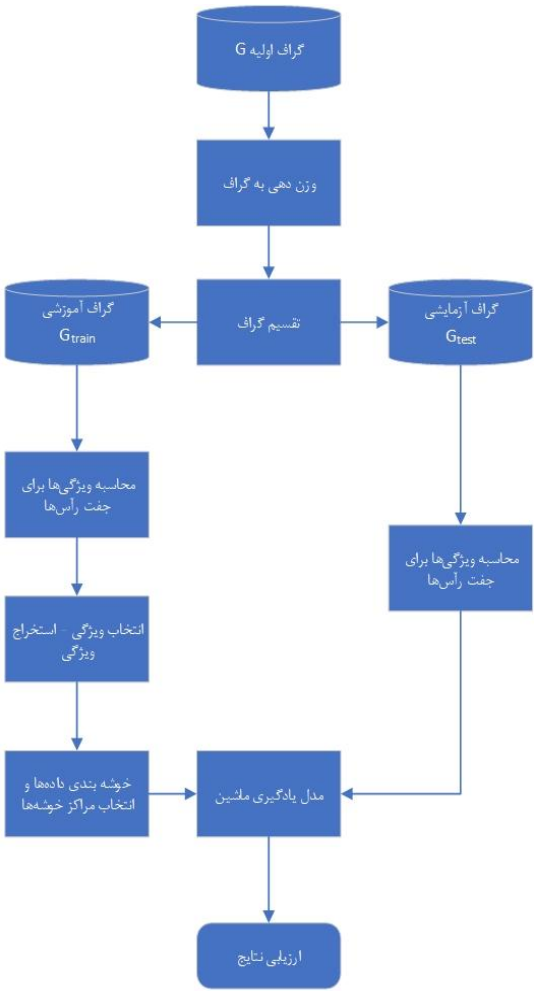
۴- روش‌های پیشنهادی

بر اساس شکل ۸، در روش‌های پیشنهادی گراف مورد پردازش ابتدا وزن‌دهی می‌شود، ابتدا ویژگی‌های ساختاری و محتوایی جدول ۲ برای جفت رأس‌های آن‌ها حساب می‌شود. سپس با استفاده از یک الگوریتم انتخاب یا استخراج ویژگی مانند PCA داده‌ها پردازش می‌شوند و ابعاد زائد حذف می‌شوند. سپس به دو قسمت آموزشی و آزمایشی تقسیم می‌شود. گراف آموزشی برای آموزش الگوریتم یادگیری ماشین استفاده می‌شود. در ادامه برای داده‌های منفی، جفت رأسهایی که در داده‌های آموزشی پیوندی بین آنها نیست، (و یا مثبت و منفی) خوشه بندی صورت می‌گیرد و مراکز خوشه‌ها به عنوان داده آموزشی برای آموزش الگوریتم یادگیری ماشین مورد استفاده قرار می‌گیرند. از سوی دیگر ویژگی‌ها برای داده‌های آزمایشی محاسبه شده و بر اساس الگوریتم آموزش داده شده پیوندهای مثبت و منفی پیش‌بینی می‌شوند و بر اساس معیار مد نظر ارزیابی می‌شوند.

جدول ۲ ویژگی‌های پیشنهادی شبکه‌های ارجاعات و همکاری نویسندگان

نوع	نام ویژگی	علامت	شبکه قابل اعمال		نوع
			ارجاعات	همکاری	
ساختاری	نزدیک‌ترین همسایه	CN	*	*	int
	آدامیک آدار	AA	*	*	float
	تخصیص منابع	RA	*	*	float
	ژاکارد	JC	*	*	float
	اتصال ترجیحی	PA	*	*	int
	نزدیک‌ترین همسایه وزن‌دار	W_CN	*	*	int
	آدامیک آدار وزن‌دار	W_AA	*	*	float
	تخصیص منابع وزن‌دار	W_RA	*	*	float
	ژاکارد وزن‌دار	W_JC	*	*	float
	اتصال ترجیحی وزن‌دار	W_PA	*	*	int
	شباهت کسینوسی بردار Node2Vec جفت رأس	N2V_COS_SIM	*	*	float
محتوایی	شباهت کسینوسی عنوان مقالات دو نویسنده	TITLE_COS_SIM		*	float
	شباهت کسینوسی چکیده دو مقاله	ABS_COS_SIM	*		float
زمانی	اختلاف سال انتشار دو مقاله	PUB_YEAR_DIFF	*		int

[1] Resource Allocation
[2] Cluster Coefficient
[3] Centrality



شکل ۸ چارچوب پیشنهادی پیش‌بینی پیوند مبتنی بر یادگیری ماشین

بر اساس روش پیشنهادی، در ادامه روش پیشنهادی برای وزن دهی شبکه، سپس ویژگی‌هایی برای جفت رأس‌های شبکه همکاری نویسندگان و ارجاعات پیشنهاد شده است. سپس روش پیش‌بینی پیوند با استفاده از یادگیری ماشین و دسته‌بندی شرح داده شده است. در انتها برای حل مشکل نمونه گیری منفی دو روش خوشه‌بندی و رگرسیون پیشنهاد شده است.

۴-۱ وزن دهی شبکه

با وجودی که در روش یادگیری ماشین هر ویژگی محتوایی را به عنوان یک درایه از بردار ویژگی در نظر می‌گیریم ولی با این حال در روش پیشنهادی از ویژگی‌های محتوایی برای وزن دهی به گراف نیز استفاده می‌کنیم. این کار به دقیق‌تر شدن نتایج شباهت‌های محلی وزن دار و در نتیجه افزایش دقت یادگیری ماشین کمک می‌کند. همچنین در صورتی که از روش یادگیری ماشین استفاده نشود، ویژگی‌های محتوایی در وزن یال‌ها در نظر گرفته شده و نادیده گرفته نخواهد شد.

با این فرضیه که در شبکه‌های علمی پیوندهای جدیدتر اهمیت بیشتری نسبت به پیوندهای قدیمی‌تر دارند و پیوندهایی با شباهت محتوایی بیشتر باید بیش از پیوندهایی با شباهت محتوایی کمتر مورد توجه قرار بگیرند، دو معیار وزن دهی برای شبکه همکاری نویسندگان (معادله ۱۲) و شبکه ارجاعات (معادله ۱۳) پیشنهاد شده است. هر دو معیار وزن‌دهی پیشنهاد شده بر اساس ترکیب خطی تأثیر زمان ایجاد پیوند و شباهت محتوایی پیوند است.

$$weight_{collab}(x,y) = \alpha\left(cosine\_sim\left(\overrightarrow{w2v(x)},\overrightarrow{w2v(y)}\right)\right) + \beta\left(\sum_{i\in collab(x,y)} \frac{1}{log(year_{current} - year_{collab}(i))}\right) \quad ۱۲$$

$$weight_{citation}(x,y) = \alpha\left(cosine\_sim\left(\overrightarrow{w2v(x)},\overrightarrow{w2v(y)}\right)\right) + \beta\left(\frac{1}{log|year_x - year_y|}\right) \quad ۱۳$$

$$\alpha,\beta > 0 \; and \; \alpha + \beta = 1$$

در هر دو معادله α و β هر دو بزرگتر از صفر و مجموع آن‌ها برابر یک است. با افزایش α و در نتیجه کاهش β تأثیر محتوا نسبت به زمان افزایش پیدا می‌کند. همچنین $\overrightarrow{w2v(x)}$ یک بردار ۱۲۸ بعدی از نهفته‌سازی محتوای متنی هر رأس است. همچنین $collab(x,y)$ مجموعه همکاری‌های دو نویسنده x و y است و $year_{collab}(i)$ سال انجام همکاری i بین این دو نویسنده است. علت استفاده از تابع لگاریتم در جریمه اختلاف سال، کم کردن تأثیر اختلاف سال زیاد ارجاع به مقاله یا همکاری است. مثلاً نسبت اختلاف ۱۰ سال و ۲۰ سال بدون استفاده از لگاریتم ۲ برابر است ولی در دنیای واقعی چندان تفاوتی بین این دو وجود ندارد، با استفاده از لگاریتم این نسبت به ۱,۳ کاهش می‌یابد.

۴-۲ تعریف ویژگی‌ها

الگوریتم‌های یادگیری ماشین برای پیش‌بینی یک برچسب یا امتیاز نیاز به داده‌های آموزشی دارند که در قالب یک بردار nبعدی به آن‌ها داده می‌شود. این ویژگی‌ها می‌توانند مبتنی بر ساختار شبکه، مبتنی بر محتوای رأس‌ها و یال‌ها و/یا مبتنی بر زمان ایجاد یک رأس یا یال باشد. با توجه به ساختار و محتوای شبکه همکاری نویسندگان و شبکه ارجاعات، مجموعه‌ای از ویژگی‌هایی که در وجود یا عدم وجود یک پیوند نقش دارد، در جدول ۲ ارائه شده است. نوع ویژگی (ستون اول) مشخص می‌کند که ویژگی عنوان شده ساختاری، محتوایی و یا زمانی است. ستون دوم جدول نام ویژگی است و ستون سوم علامت اختصاری است که در نتایج آزمایشها از آن استفاده خواهد شد. از آنجا که در شبکه همکاری نویسندگان متن مقاله وجود ندارد، برخی ویژگی‌ها در این شبکه قابل استفاده نیست؛ مانند شباهت محتوایی دو مقاله. بنابراین در شبکه همکاری نویسندگان تنها از شباهت میان عنوان مقاله استفاده شده است. آخرین ستون جدول ۲، نوع هر ویژگی را نشان می‌دهد؛ ویژگیهایی که بازه مقداری برای آنها مشخص نشده است می‌توانند هر مقداری بزرگتر از صفر داشته باشند. ویژگیهایی که در جدول ۲ آمده است، ویژگیهایی هستند که محاسبه آنها از نظر زمان اجرا مقیاس‌پذیر است. ویژگیهای دیگری مانند معیارهای مرکزیت وجود دارند اما با افزایش اندازه گراف زمان محاسبه آنها به صورت نمایی افزایش می‌یابد. به همین دلیل چنین ویژگی‌هایی در اینجا استفاده نشده است.

۴-۳ پیش‌بینی پیوند با استفاده از یادگیری ماشین

گراف $G=(V,E)$ به طوری که V مجموعه رأس‌های گراف؛ در اینجا نویسندگان یا مقالات و E پیوندهای موجود میان رأس‌های شبکه می‌باشد. مجموعه E'، مجموعه پیوندهایی است که در گراف وجود ندارد ولی به صورت بالقوه امکان ایجاد در آینده را دارد. برای پیش‌بینی پیوند با استفاده از یادگیری ماشین مدل M آموزش داده می‌شود به طوری که هر پیوند عضو E را به یک احتمال بین صفر و یک نگاشت می‌کند. مدل M در این روش پیشنهادی می‌تواند یک الگوریتم دسته‌بندی یا رگرسیون باشد.

برای حل مسئله‌ی پیش‌بینی پیوند با استفاده از یادگیری ماشین، گراف باید در قالب مجموعه‌ای از بردارها بازنمایی شود. این بردارها برای هر جفت رأس محاسبه می‌شوند درایه‌های این بردار می‌تواند ویژگی‌های ساختاری یا محتوایی مربوط به جفت رأس باشد. مجموعه‌ای از ویژگی‌های قابل محاسبه برای بردار ویژگی‌ها در جدول ۲ آورده شده است.

بعد از آنکه برای هر جفت رأس، بردار ویژگی حساب شد، یک برچسب وجود یا عدم وجود پیوند برای آن مشخص می‌شود. مدل یادگیری ماشین این برچسب را برای جفت رأس‌های جدید پیش‌بینی می‌کند. به طور خلاصه مراحل در زیر بیان شده است:

۱ ویژگی‌هایی برای بردارها تعریف می‌شود (مانند معیارهای شباهت، میزان مشابهت محتوایی و زمان ایجاد پیوند).

۲ به ازای تمام جفت رأس‌های ممکن در گراف آموزشی بردار ایجاد و مقادیر ویژگی‌های هر جفت رأس محاسبه می‌شود.

۳ با توجه به این که جفت رأس تشکیل دهنده بردار با یکدیگر پیوند دارند یا خیر، برچسب «وجود پیوند» یا «عدم وجود پیوند» انتصاب داده می‌شود.

۴ الگوریتم یادگیری ماشین با استفاده از بردارهای به دست‌آمده آموزش داده می‌شود.

۵ مدل نهایی برای هر جفت رأس جدید در گراف، پیش‌بینی پیوند می‌کند.

به عنوان مثال در جدول ۳ برای هر جفت رأس در شبکه یک بردار ویژگی (یک سطر جدول) بدست آمده است. الگوریتم پس از یادگیری این داده‌ها می‌تواند برچسب یک داده جدید را پیش‌بینی کند.

جدول ۳ مثالی از بردارهای ورودی الگوریتم پیش‌بینی پیوند (مجموعه آموزشی) √: وجود پیوند و ×: عدم وجود پیوند

جفت رأس	JC	N2V_COS_SIM	TITLE_COS_SIM	...	برچسب
A - B	۰,۶۷	۰,۴۰	۰,۵۱	...	√
A - C	۰,۲۳	۰,۵۹	۰,۳۱	...	×
B - C	۰,۶۳	۰,۲۲	۰,۳۹	...	×
A - D	۰,۹۵	۰,۸۵	۰,۷۱	...	√
...	...	...	...	...	...

۴-۴ حل مشکل نمونه‌گیری منفی با خوشه‌بندی

گراف‌های دنیای واقعی معمولاً پراکنده[1] هستند؛ یعنی تعداد پیوندهایی که وجود دارند بسیار بیشتر از تعداد پیوندهایی هستند که وجود ندارند. مثلا در گراف همکاری نویسندگان ogbl-collab که روش پیشنهادی روی آن آزمایش خواهد شد، میانگین درجه‌ی رئوس ۸ و تعداد رئوس ۲۳۵،۸۶۸ است. یعنی هر نویسنده به طور میانگین با ۸ نویسنده همکاری کرده و با ۲۳۵،۸۶۰ نویسنده همکاری نکرده است.

از طرفی برای آموزش مدل یادگیری ماشین به تعدادی نمونه‌ی منفی و تعدادی نمونه‌ی مثبت نیاز است. با توجه به پراکنده بودن گراف مورد پردازش، می‌توان تمام نمونه‌های مثبت را به عنوان داده‌ی آموزشی به مدل یادگیری ماشین در نظر گرفت ولی تعداد نمونه‌های منفی در این گراف بسیار زیاد است. به عنوان مثال، در مجموعه داده ogbl-collab تعداد ۲۳۵،۸۶۸ رأس داریم که بر اساس معادله ۱۴ تعداد رأس‌های گراف کامل، تقریبا ۲۸ میلیارد یال ممکن وجود دارد که از این بین ۱،۲۸۵،۴۶۵ یال مثبت و سایر یال‌ها منفی هستند.

$$complete\_graph\_edges = \frac{n \times (n-1)}{2} = \frac{235{,}868 * 235{,}867}{2} \cong 28 \times 10^9 \qquad ۱۴$$

محاسبه ویژگی‌ها برای این تعداد داده‌ی آموزشی منفی بسیار زمان‌بر است و نیاز به منابع سخت‌افزاری زیادی دارد، همچنین الگوریتم‌های یادگیری ماشین معمول (غیر از آن‌هایی که برای داده‌های حجیم پیاده سازی شدند) برای پردازش این تعداد داده نیاز به زمان و منابع سخت‌افزاری زیادی دارند. از طرف دیگر، بسیاری از این داده‌ها تکراری هستند و اطلاعات تازه‌ای در اختیار

[1] Sparse

الگوریتم برای تشخیص پیوندهای منفی قرار نمی‌دهند و حتی ممکن است باعث بیش برازش مدل یادگیری ماشین نیز بشود.

برای کم کردن تعداد نمونه‌های منفی می‌توان از میان نمونه‌های منفی، تعدادی را به صورت تصادفی انتخاب کرد. با انتخاب تصادفی ممکن است برخی از نمونه‌هایی که در آموزش بهتر مدل می‌توانند تاثیرگزار باشند انتخاب نشوند و از طرفی داده‌هایی که شبیه به یکدیگر هستند چندین بار انتخاب شوند.

برای حل این مشکل دو روش مبتنی بر خوشه‌بندی و رگرسیون پیشنهاد شده است. در روش مبتنی بر خوشه‌بندی ابتدا نمونه‌های منفی به K خوشه تقسیم و از هر خوشه یک نماینده انتخاب می‌شود. به این ترتیب از تمام فضای نمونه‌های منفی به طور یکنواخت نمونه گیری شده و نمونه‌های منفی با یکدیگر متفاوت خواهند بود. با این روش تعداد نمونه‌های منفی کاهش و کیفیت آن‌ها افزایش پیدا خواهد کرد. به این ترتیب می‌توان با منابع سخت‌افزاری در دسترس و الگوریتم‌های یادگیری ماشین رایج مسئله دسته‌بندی را حل کرد. پیش‌بینی می‌شود استفاده از همین روش برای داده‌های مثبت نیز تأثیر مثبتی داشته باشد. در هر گراف مورد پردازش ممکن است مقدار بهینه برای K متفاوت باشد. پیشنهاد ما برای یافتن K بهینه استفاده از آزمون و خطا و یا معیارهای ارزیابی خوشه‌بندی است.

۵- آزمایش‌ها

در این بخش پنج آزمایش برای مقایسه روشهای پیشنهادی و سایر روشهای موجود بر اساس معیارهای ارزیابی عملکرد HITS و MRR بر روی دو مجموعه داده ارجاعات و همکاری نویسندگان طراحی شده است. ابتدا ابزارهای نرم‌افزاری و سخت‌افزاری شرح داده می‌شوند. سپس مجموعه داده‌های مورد استفاده و معیارهای ارزیابی مورد بررسی قرار می‌گیرند و در ادامه برای هر آزمایش، شرح آزمایش، نتایج و تحلیل ارائه می‌شود.

۵-۱ ابزارهای مورد استفاده

سخت‌افزار مورد استفاده برای آزمایشات این فصل یک رایانه‌ی خانگی با مشخصات Intel Pentium Silver N5000 2.7GHZ 8GB Ram بوده است. البته برخی آزمایشات مربوط به روش‌های جاسازی با الگوریتم Node2Vec به دلیل زمان زیاد اجرا روی این سیستم قابل اجرا نبودند. این آزمایشات روی سرور محاسبات ابری دانشگاه فردوسی مشهد[1] با پردازنده ۳۲ هسته‌ای و ۶۴ گیگابایت رم اجرا شد.

آزمایشات با زبان پایتون و در محیط Jupyter-Lab پیاده سازی شدند. برای پردازش‌های مربوط به گراف از NetworkX [۲۹] و برای پیاده‌سازی الگوریتم‌های مربوط به یادگیری ماشین و انتخاب و استخراج ویژگی از scikit-learn استفاده شد. برای بارگزاری داده‌ها، تقسیم گراف به مجموعه‌ی آموزشی، اعتبارسنجی و آزمایشی و در نهایت ارزیابی و مقایسه نتایج از چارچوب OGB[2] استفاده کردیم. توضیحات این چارجوب در پیوست ب ارایه شده است.

۵-۲ مجموعه داده‌ها

در این بخش از دو مجموعه داده OGB استفاده می‌شود که در زمینه پیش‌بینی پیوند و شبکه‌های علمی هستند. مجموعه داده ogbl-collab یک شبکه از نویسندگان و همکاری‌های بین آن‌هاست. عنوان مقالات هر نویسنده در رأس‌ها و سال همکاری

[1] https://ferdowsi.cloud/

[2] Open Graph Benchmark

در یال‌ها به عنوان فرا داده در مجموعه داده موجود است. مجموعه داده ogbl-citation2 نیز یک شبکه ارجاعات است که متشکل از مقالات و ارجاعات بین آن‌ها است. در رأس‌های این شبکه سال انتشار مقاله و عنوان چکیده مقاله به عنوان ویژگی وجود دارد. هر دو این مجموعه داده‌ها بر اساس زمان به دو مجموعه آموزشی و آزمایشی تقسیم می‌شود.

جدول ۴ مشخصات مجموعه‌داده‌های مربوط به شبکه‌های علمی در OGB

	ogbl-collab	ogbl-citation2
مقیاس	متوسط	بزرگ
# رأس‌ها	۲۳۵،۸۶۸	۲،۹۲۷،۹۶۳
# یال‌ها	۱،۲۸۵،۴۶۵	۳۰،۵۶۱،۱۸۷
تقسیم بر اساس	زمان	زمان
داده محتوایی	عنوان مقالات نویسنده، سال همکاری	عنوان و چکیده مقاله، سال انتشار
معیار ارزیابی	Hits@50	MRR

۵-۳ معیارهای ارزیابی

برای پیش‌بینی پیوند معیارهای ارزیابی زیادی وجود دارد و همچنین می‌توان از معیارهای ارزیابی دسته‌بندی برای پیش‌بینی پیوند استفاده کرد. ولی با توجه به این که در آزمایشات از چارچوب OGB استفاده می‌شود، در انتخاب معیار ارزیابی قدرت انتخاب وجود ندارد. OGB برای مجموعه داده ogbl-collab از معیار Hits@50 و برای ogbl-citation2 از معیار MRR[1] استفاده می‌کند. در ادامه هر دو معیار شرح داده می‌شوند.

۵-۳-۱ معیار Hits@k

معیار Hits@k نشان می‌دهد چه نسبتی از پیوندهای آزمایشی مثبت در میان مجموعه‌ای از پیوندهای منفی پس از مرتب‌سازی نزولی میان k پیوند اول قرار می‌گیرد. در **این** معیار هر چقدر k عدد کوچکتری باشد معیار سختگیرانه تر می‌شود. این معیار در معادله ۱۵ نشان داده شده است. در این معادله t داده مورد آزمایش، $\boldsymbol{S_{test}}$ مجموعه داده‌های آزمایشی و $rank(t)$ رتبه پیوند مثبت در مجموعه آزمایشی است. هر چقدر الگوریتم پیش‌بینی پیوند بتواند امتیازی بالاتری به پیوند مثبت نسبت به پیوندهای منفی انتصاب دهد، پیوند مثبت در رتبه‌ی بالاتری قرار می‌گیرد و مقدار معیار Hits افزایش می‌یابد (برای اطلاعات بیشتر به پیوست ب مراجعه فرمایید).

$$\boldsymbol{Hits@k = \frac{|\{t \epsilon S_{test} | rank(t) \leq k\}|}{|S_{test}|}} \qquad ۱۵$$

۵-۳-۲ معیار MRR

میانگین رتبه متقابل یا MRR یک امتیاز نسبی است که میانگین یا میانگین معکوس رتبه‌هایی را که در آن اولین سند مرتبط برای مجموعه‌ای از جستارها بازیابی شده است. این معیار در اصل مربوط به بازیابی اطلاعات و جستجو است ولی در پیش‌بینی پیوند به معنای میانگین رتبه یک پیوند مثبت در میان مجموعه‌ای پیوندهای منفی است. این معیار با معادله ۱۶ نشان داده می‌شود. در این معادله Q مجموعه پیوندهای منفی به علاوه پیوند مثبت مد نظر و $rank(t)$ به معنای رتبه پیوند مثبت در میان پیوندهای منفی پس از مرتب سازی نزولی است. هر چقدر مدل پیش بینی کننده به پیوند مثبت امتیاز بیشتری نسبت به پیوندهای منفی بدهد و رتبه بالاتری کسب کند مقدار MRR بیشتر می‌شود.

$$\boldsymbol{MRR = \frac{1}{|Q|} \sum_{t=1}^{|Q|} \frac{1}{rank(t)}} \qquad ۱۶$$

۵-۴ آزمایش اول: معیارهای شباهت محلی

[1] Mean Reciprocal Rank

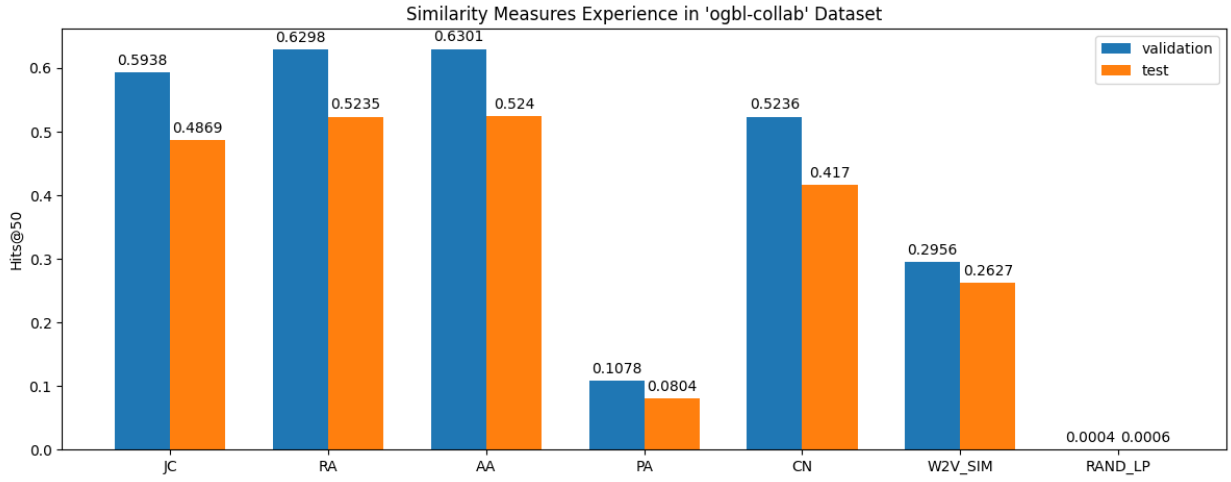
در این بخش، یک آزمایش برای مقایسه روشهای مبتنی بر شباهت محلی گراف طراحی شده است. در این آزمایش و آزمایش‌های بعدی داده مورد آزمایش ogbl-collab و ogbl-citation2 خواهد بود که به ترتیب بر اساس معیارهای Hits@50 و MRR ارزیابی می‌شوند. همچنین این آزمایش با یک پیش‌بینی کننده تصادفی با نام random_lp تکرار شده که یک عدد تصادفی بین ۰ و ۱ به هر جفت رأس انتصاب می‌دهد. معیار random_lp را برای این به آزمایش اضافه کردیم که ببینیم معیارهای شباهت محلی نسبت به یک روش تصادفی چقدر بهتر عمل می‌کنند. همچنین از شباهت محتوایی W2V_SIM نیز استفاده کردیم که فارغ از ساختار شبکه، صرفاً شباهت کسینوسی بردار نهفته‌سازی دو رأس را محاسبه می‌کند.

همانطور در شکل ۹ و شکل ۱۰ مشخص است معیار آدامیک آدار (AA) در هر دو مجموعه داده و پس از آن تخصیص منبع (RA) بهترین نتایج را کسب کرده است. به نظر می‌رسد این دو معیار به دلیل جریمه کردن درجه همسایگان مشترک توانستند نسبت به بقیه نتیجه بهتری کسب کنند. جریمه کردن درجه همسایگان مشترک در کاربردهای دیگر مانند شبکه‌های اجتماعی تأثیر مثبتی داشته و می‌بینیم در شبکه‌های علمی نیز این قاعده برقرار است. یعنی مقالاتی که ارجاعات زیادی دارند و نویسندگانی که همکاری‌های زیادی انجام دادند نسبت به مقالات کم ارجاع و نویسندگان با تعداد کم همکاری، همسایه‌های کم اهمیت تری هستند. همچنین معیار اتصال ترجیحی (PA) بدترین نتایج را کسب کرده است ولی با این حال با اختلاف زیادی از پیش‌بینی کننده‌ی تصادفی بهتر عمل کرده است. شباهت کسینوسی بردار نهفته‌سازی شده دو رأس (W2V_SIM) بدون در نظر گرفتن ساختار گراف، در هر دو مجموعه داده از پیش‌بینی کننده تصادفی بهتر عمل کرده است. این موضوع نشان می‌دهد محتوای متنی رأس‌های شبکه علمی می‌توانند در پیش‌بینی پیوند تأثیر گذار باشند. شباهت محتوا در شبکه‌ی همکاری نویسندگان بهتر از شبکه ارجاعات عمل کرده است، از این موضوع نتیجه می‌گیریم همکاری که بین دو نویسنده شکل می‌گیرد در مقایسه با ارجاع مقالات، وابستگی بیشتر به موضوع و محتوا دارد.

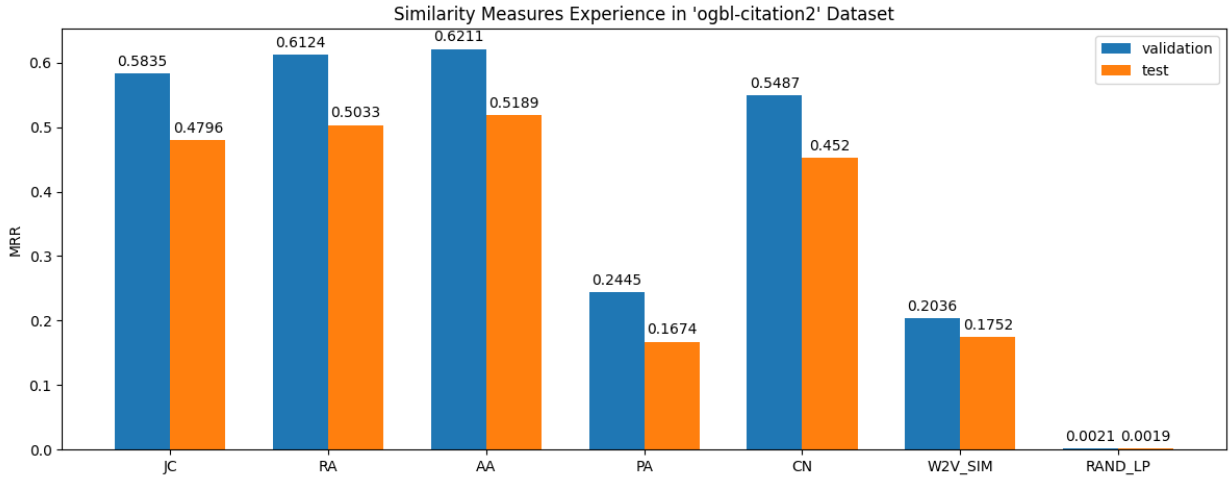
نتایج مدل تصادفی را با استفاده از احتمال می‌توان توجیه کرد. در ogbl-collab تعداد پیوندهای منفی ساختگی ۱۰۰،۰۰۰ و پیوند مثبت واقعی در Hits@50 باید در بین ۵۰ داده اول رتبه‌بندی شود. بر اساس معادله ۱۷، نتیجه به دست آمده از طریق احتمال با نتیجه حاصل از آزمایش تفاوت زیادی ندارد. همچنین در ogbl-citation2 میانگین معکوس رتبه پیوند هدف در میان ۱۰۰۰ رأس منفی به عنوان MRR محاسبه می‌شود. نتیجه پیش‌بینی تصادفی در معادله ۱۸ با نتایج آزمایش مطابقت دارد. با افزایش تعداد آزمایش‌ها و میانگین گرفتن از نتایج، عدد میانگین به عدد به دست آمده از طریق احتمال همگرا می‌شود.

$$Hits@50_{rand} = \frac{50}{100{,}000} = 0.0005 \qquad ۱۷$$

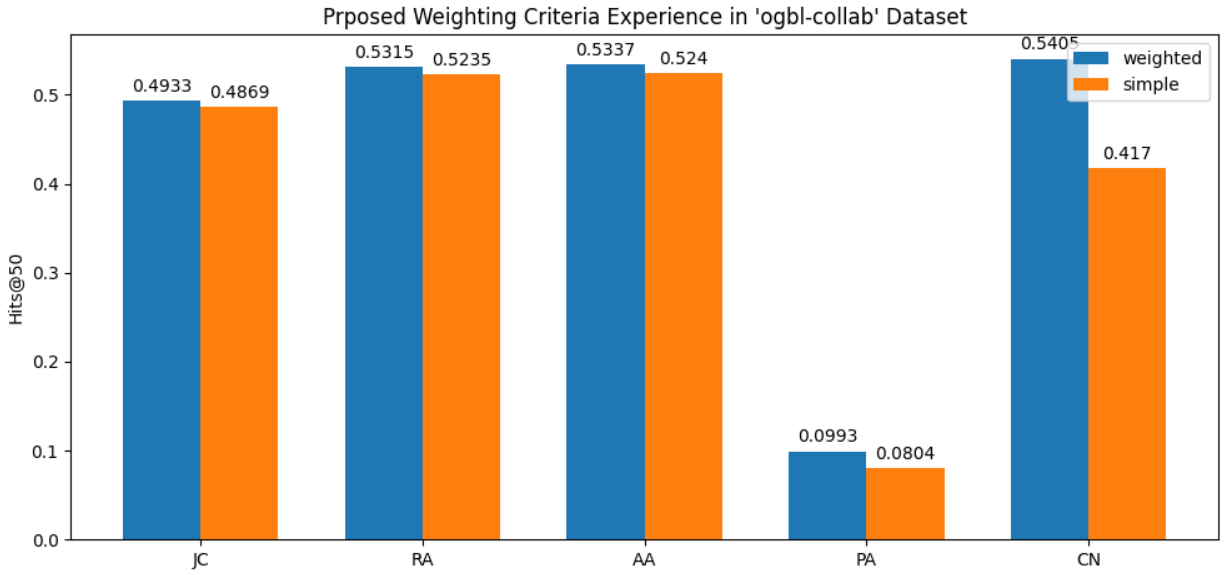
$$MRR_{rand} = \frac{1}{1000}\sum_{i=1}^{1000}\frac{1}{500} = 0.002 \qquad ۱۸$$



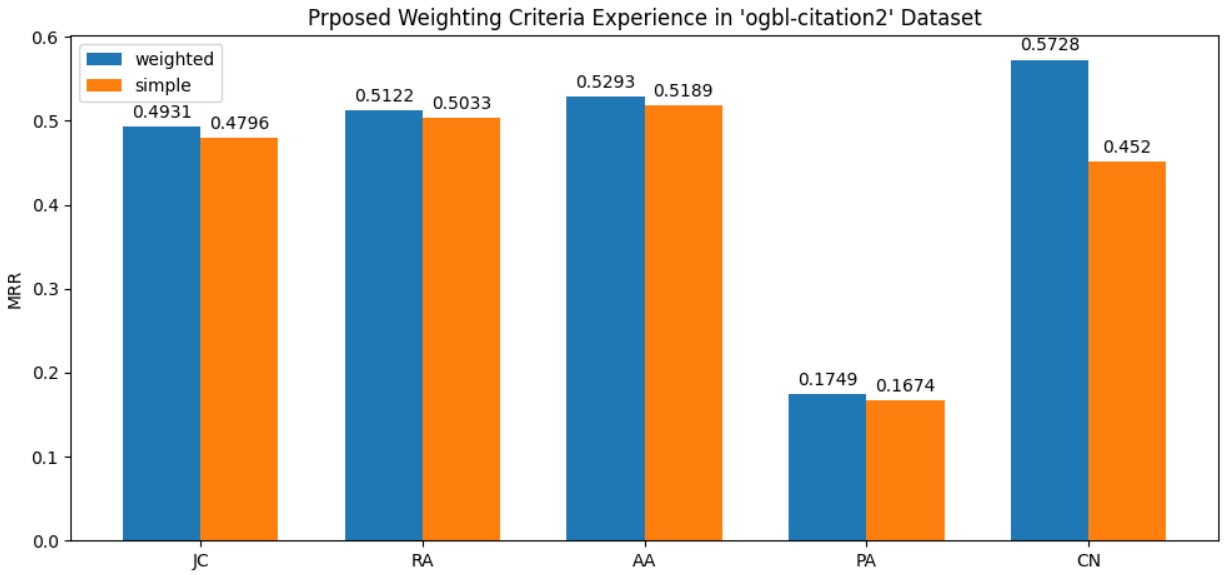
شکل ۹ مقادیر Hits@50 بر اساس استفاده از معیارهای شباهت محلی در مجموعه داده ogbl-collab



شکل ۱۰ مقادیر MRR بر اساس استفاده از معیارهای شباهت محلی در مجموعه داده ogbl-citation2



شکل ۱۱ مقادیر Hits@50 بر اساس استفاده از روش پیشنهادی وزن دهی و مقایسه با روش بدون وزن در داده ogbl-collab



شکل ۱۲ مقادیر MRR بر اساس استفاده از روش پیشنهادی وزن دهی و مقایسه با روش بدون وزن در داده ogbl-citation2

۵-۵ آزمایش دوم: روش وزن دهی پیشنهادی

این آزمایش برای بررسی روش‌های وزن دهی پیشنهاد شده در معادله ۱۲ و معادله ۱۳ طراحی شده است. در این آزمایش شبکه‌های همکاری نویسندگان و ارجاعات بر اساس روش پیشنهادی وزن دار شدند و با استفاده از معیارهای شباهت وزن دار که در مفاهیم پایه شرح داده شدند مورد پردازش قرار گرفتند. سپس نتایج هر معیار با نسخه بدون وزن (آزمایش اول) مقایسه

می‌شود. در این آزمایش و آزمایش‌های بعدی فقط داده‌های آزمایشی در نتایج آورده می‌شوند و داده‌های اعتبار سنجی را برای جلوگیری از نمودارهای اضافی و آسان‌تر شدن مقایسه نتایج گزارش نمی‌کنیم. مقدار بهینه متغیرهای α و β برای این آزمایش که با آزمون و خطا به دست آمدند در جدول ۵ آمده است. نتایج این آزمایش در مقایسه با آزمایش قبل نیز در شکل ۱۱ و شکل ۱۲ قابل مشاهده است.

جدول ۵ مقدار α و β در شبکه همکاری نویسندگان و ارجاعات

شبکه	متغیر	مقدار
همکاری نویسندگان(معادله ۱۲)	α (تاثیر محتوا)	0.5
	β (تاثیر زمان)	0.5
ارجاعات(معادله ۱۳)	α (تاثیر محتوا)	0.4
	β (تاثیر زمان)	0.6

نتایج نشان می‌دهد تمام معیارهای برای هر مجموعه داده تقریباً به میزان ثابتی بهبود پیدا کردند. البته معیار همسایه مشترک (CN) در هر دو مجموعه داده حدود ٪۱۴ افزایش داشته. همسایه مشترک پایه معیارهای دیگر مانند ژاکارد و آدامیک آدار و تخصیص منبع است منتها نسبت به آن‌ها رفتار خیلی ساده تری دارد. احتمالا به همین دلیل رفتار ساده (تعداد همسایه‌های مشترک) در مقایسه با آدامیک آدار و تخصیص منبع که همسایگان مشترک را بر اساس درجه جریمه می‌کنند یا ژاکارد که نسبت همسایه‌های مشترک به تمام همسایه‌ها را به عنوان شباهت محاسبه می‌کند عملکرد ضعیف تری دارد. با معیار وزن دهی پیشنهاد شده، ارزش هر همسایه مشترک تفاوت خواهد کرد. همسایه مشترک وزن‌دار مجموع ارزش همسایه‌ها را عنوان شباهت محاسبه می‌کند و به همین خاطر در مقایسه با همسایه مشترک ساده بسیار هوشمندانه تر عمل می‌کند و توانسته نتایجی مشابه سایر معیارها (به غیر از اتصال ترجیحی) کسب کند.

همچنین جدول ۵ نشان می‌دهد در شبکه همکاری نویسندگان تاثیر زمان و محتوا به یک اندازه است. ولی در شبکه ارجاعات تأثیر زمان بیشتر از محتوا است؛ به عبارت دیگر تمایل ایجاد پیوند در شبکه ارجاعات بیشتر با رأس‌های جدیدتر است تا رأس‌های با شباهت محتوایی بیشتر.

۵-۶ آزمایش سوم: پیش‌بینی پیوند با SVM و نمونه‌گیری منفی تصادفی

این آزمایش برای بررسی روش‌های پیشین مبتنی بر تک ویژگی (شباهت محلی وزن‌دار و بدون وزن) با روش یادگیری ماشین طراحی شده است. در این آزمایش، به تعداد یال‌های آموزشی با برچسب مثبت، یال‌هایی با برچسب منفی به طور تصادفی انتخاب کردیم، و ویژگی‌های ذکر شده در جدول ۲ را برای آن‌ها محاسبه کردیم. به این ترتیب هر پیوند مثبت یا منفی را به یک بردار ۱۳بعدی تبدیل کردیم. با توجه به این که مقادیر ویژگی‌های محاسبه شده در بازه‌های متفاوتی قرار دارند و بدون نرمال سازی باعث می‌شود برخی الگوریتم‌های یادگیری ماشین به سمت ویژگی‌هایی با مقدار بزرگتر سوگیری کنند. به همین خاطر مقادیر را با تابع StandardScaler از بسته scikit-learn نرمال کردیم.

در این آزمایش از SVM در بسته scikit-learn با هسته linear به عنوان الگوریتم یادگیری ماشین استفاده شده است. با توجه به این که معیارهای MRR و Hits@k هر دو بر اساس مرتب سازی بر حسب امتیاز عمل می‌کنند و بر خلاف مسئله دسته‌بندی، داده‌های آزمایشی بر حسب برچسب مثبت و منفی مطلق مورد ارزیابی قرار نمی‌گیرد و احتمال وجود پیوند در ارزیابی استفاده می‌شود، پارامتر probability را برابر true قرار دادیم که به

جای یک برچسب، احتمال تعلق به آن برچسب (وجود یا عدم وجود پیوند) را به عنوان خروجی محاسبه کند.

در این آزمایش با توجه به انتخاب تصادفی داده‌های منفی، ممکن است نتیجه با هر بار اجرا تغییر کند، به همین دلیل این آزمایش را پنج بار با داده‌های تصادفی متفاوت تکرار کردیم و نتایج را در جدول ۶ نشان دادیم.

جدول ۶ نتایج پیش‌بینی پیوند با SVM و نمونه‌گیری منفی تصادفی

	تکرار ۱	تکرار ۲	تکرار ۳	تکرار ۴	تکرار ۵	میانگین
Hits@50	0.513	0.503	0.528	0.508	0.503	0.511 ± 0.009
MRR	0.490	0.506	0.493	0.509	0.503	0.500 ± 0.007

با توجه به اختلاف نتایج در هر بار تکرار، به نظر می‌رسد داده‌های آموزشی در کیفیت پیش‌بینی پیوند تأثیر داشته باشد و به دلیل تصادفی انتخاب شدن در هر تکرار، نتایج با تکرارهای قبل تفاوت دارد. به طور کلی نتایج این آزمایش نسبت به بهترین شباهت‌های وزن دار آزمایش قبل بهبود خاصی نداشته. دلیل این موضوع می‌تواند ناکافی بودن داده‌های آموزشی و انتخاب تصادفی این داده‌ها باشد. همچنین ممکن است وجود ویژگی‌هایی که با سایر ویژگی‌ها همبستگی دارند نیز یکی از علت‌های نتیجه نامطلوب باشد. این موارد در آزمایش‌های بعدی بررسی خواهند شد.

۵-۷ آزمایش چهارم: استخراج ویژگی با PCA

از آنجایی که برخی ویژگی‌های جدول ۲ ممکن است با یگدیگر همبستگی داشته باشند، این آزمایش را طراحی کردیم که با استفاده از الگوریتم استخراج ویژگی PCA ابعاد داده‌ها را کاهش دهیم در این آزمایش بهترین نتایج با n_component=5 حاصل شد، این مقدار با آزمون و خطا به دست آمد. مانند آموزش قبل با توجه به تصادفی انتخاب شدن داده‌های آموزشی، به دلیل امکان اختلاف نتایج آزمایش را تکرار کردیم و نتایج هر تکرار را در جدول ۷ گزارش کردیم.

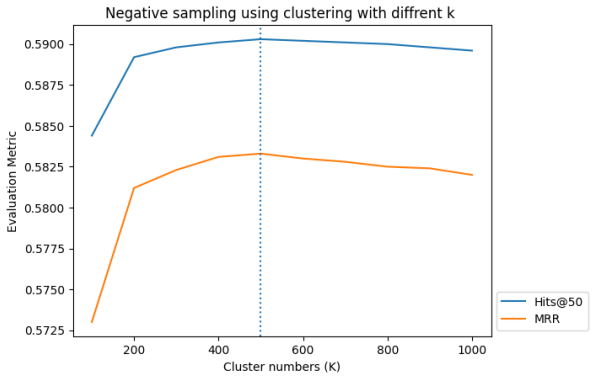
جدول ۷ نتایج پیش‌بینی پیوند با PCA و SVM

SVM	تکرار ۱	تکرار ۲	تکرار ۳	تکرار ۴	تکرار ۵	میانگین
Hits@50	0.540	0.536	0.531	0.545	0.537	0.538 ± 0.004
MRR	0.525	0.528	0.529	0.526	0.526	0.527 ± 0.001

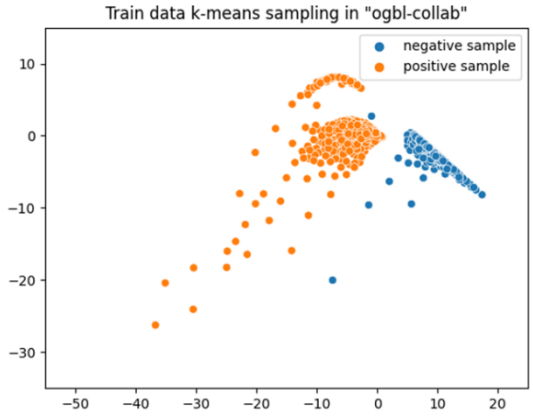
همان طور که انتظار می‌رفت، مشاهده می‌شود نتایج نسبت به آزمایش پیش که در شرایط یکسان و بدون PCA انجام شد بهبود پیدا کرده و مانند بسیاری از مسائل یادگیری ماشین PCA باعث افزایش دقت مدل یادگیری ماشین شده. با توجه به این تأثیر مثبت، در آزمایشات بعدی که از خوشه‌بندی و رگرسیون استفاده خواهد شد، از ویژگی‌های استخراج شده توسط PCA استفاده می‌شود.

۵-۸ آزمایش پنجم: نمونه‌گیری با خوشه‌بندی

این آزمایش برای بررسی نتایج روش پیشنهادی برای نمونه‌گیری منفی با استفاده از خوشه بندی طراحی شده است. به دلیل تعداد زیاد داده الگوریتم‌های خوشه‌بندی معمولی زمان زیادی برای اجرا نیاز دارند به همین خاطر در این آزمایش از الگوریتم MiniBatchKMeans با batch_size=1024 استفاده شده است. داده‌های مثبت و منفی پس از استخراج ویژگی با PCA خوشه بندی شدند و الگوریتم یادگیری ماشین بر اساس مرکز هر خوشه آموزش داده شد. برای این کار آزمایش را برای k=100 تا k=1000 تکرار کردیم و نتایج را در شکل ۱۳ نشان دادیم. همچنین برای درک بهتر روش پیشنهادی یک بار ۵۰۰ داده مثبت و منفی را از مجموعه داده ogbl-collab یک بار به صورت تصادفی و بار دیگر با استفاده از خوشه بندی انتخاب کردیم و داده‌ها را به صورت دو بعدی در شکل ۱۴ نشان دادیم.



شکل ۱۳ ارزیابی نتایج پیش‌بینی پیوند با نمونه‌گیری منفی با خوشه بندی به ازای تعداد خوشه‌های متفاوت



شکل ۱۴ مقایسه نمونه‌گیری تصادفی و استفاده از خوشه‌بندی در داده‌های آموزشی

در شکل ۱۴ سمت چپ یک بار از داده‌های آموزشی به صورت تصادفی ۵۰۰ داده مثبت و منفی انتخاب کردیم، و یکبار همین داده‌ها را به ۵۰۰ خوشه تقسیم کردیم و مراکز خوشه‌ها را در سمت راست نشان دادیم. با وجود این که در هر دو سمت شکل ۵۰۰ داده مثبت و منفی انتخاب شده ولی داده‌های انتخابی در نمونه‌گیری تصادفی بسیار نزدیک به هم تر و متراکم تر نسبت به روش خوشه‌بندی هستند که نشان می‌دهد بسیاری از داده‌ها تقریباً تکراری هستند. هنگامی که الگوریتم با نمونه گیری تصادفی آموزش داده شود بسیاری از حالات مثبت یا منفی که در داده‌های اصلی وجود دارد آموزش داده نخواهد شد. در صورتی که نمونه گیری با خوشه بندی حالات بیشتری از هر برچسب مثبت یا منفی را پوشش می‌دهد.

۶- نتیجه‌گیری

شبکه‌های علمی مانند شبکه همکاری نویسندگان و شبکه ارجاعات شبکه‌هایی هستند که نحوه تکامل آن‌ها در آینده قابل پیش‌بینی است. با پیش‌بینی این تکامل، می‌توان برای یافتن متخصص و منابع مناسب برای پژوهش‌های آتی کمک گرفت. در این مقاله یک چارچوب کلی برای مسائل پیش‌بینی پیوند در شبکه‌های علمی پیشنهاد شد که اساس آن ترکیب ویژگی‌های محتوایی و ساختاری در قالب یک بردار ویژگی و استفاده از الگوریتم‌های یادگیری ماشین است؛ برای کاهش ابعاد از PCA استفاده شد. با توجه به حجم زیاد داده‌ها در شبکه‌های علمی دنیای واقعی و پراکنده بودن پیوندهای شبکه، نمونه‌های منفی بسیار بیشتر از نمونه‌های مثبت هستند که با روش پیشنهادی در این مقاله، استفاده از خوشه بندی داده‌های منفی و انتخاب مراکز خوشه‌ها، نمونه‌های محدودی به عنوان نماینده انتخاب شد که با این روش نیز نتایج بار دیگر بهبود یافت.

پیاده‌سازی روش‌های یادگیری ماشین روی شبکه‌های علمی با توجه به برخی ویژگی‌های آن مفید باشد. به عنوان مثال توجه به حوزه‌های تحقیقاتی که می‌توانند در گرافهای علمی به صورت زیرگراف (یا جامعه) در نظر گرفته شود. همچنین آزمایش بیشتر روی یافتن ترکیب ماسبتری از محتوا و ساختار و شاید مشارکت سایر ویژگیهای شبکه‌های علمی می‌تواند یکی از کارهای آینده در این حوزه باشد.

مراجع

[1]. R. Taimourei-Yansary, M. Mirzarezaee, M. Sadeghi, and B. N. Araabi, "Predicting invasive disease-free survival time in breast cancer patients using semi-supervised graph-based machine learning techniques", *Soft Computing Journal*, vol. 10, no. 1, pp. 48-69, 2022.

[2]. E. Mahfooz and G. Fath-Tabar, "Sum of distance between vertices of graphs", *Soft Computing Journal*, vol. 5, no. 2, pp. 28-33, 2021.

[3]. A. Keypour, "Link Prediction in Social Networks through classifiers combination", *Soft Computing Journal*, vol. 4, no. 2, pp. 2-17, 2021.

[4]. V. Martínez, F. Berzal, and J.-C. Cubero, "A survey of link prediction in complex networks", *ACM computing surveys (CSUR),* vol. 49, no. 4, pp. 1-33, 2016, DOI: https://doi.org/10.1145/3012704.

[5]. L. Lü, and T. Zhou, "Link prediction in complex networks: A survey," *Physica A: statistical mechanics and its applications,* vol. 390, no. 6, pp. 1150-1170, 2011, https://doi.org/10.1016/j.physa.2010.11.027

[6]. C. P. Muniz, R. Goldschmidt, and R. Choren, "Combining contextual, temporal and topological information for unsupervised link prediction in social networks," *Knowledge-Based Systems,* vol. 156, pp. 129-137, 2018, https://doi.org/10.1016/j.knosys.2018.05.027.

[7]. M. Nikkar, R. Alijani, and K. M. H. GHAZIZADEH, "Investigation of the presence of surgery researchers in research gate scientific network: An altmetrics study," *Iranian Journal of Surgery,* vol. 25, no. 2, pp. 76-82, 2017.

[8]. H. Liu, H. Kou, C. Yan, and L. Qi, "Link prediction in paper citation network to construct paper correlation graph," *EURASIP Journal on Wireless Communications and Networking,* vol. 2019, no. 1, pp. 1-12, 2019.

[9]. E. Bütün, and M. Kaya, "Predicting citation count of scientists as a link prediction problem," *IEEE transactions on cybernetics,* vol. 50, no. 10, pp. 4518 - 4529, 2019, https://doi.org/10.1109/TCYB.2019.2900495.

[10]. N. Shibata, Y. Kajikawa, and I. Sakata, "Link prediction in citation networks," *Journal of the American society for information science and technology,* vol. 63, no. 1, pp. 78-85, 2012, https://doi.org/10.1002/asi.21664.

[11]. V. Latora, V. Nicosia, and G. Russo, *Complex networks: principles, methods and applications*: Cambridge University Press, 2017.

[12]. P. M. Chuan, M. Ali, T. D. Khang, and N. Dey, "Link prediction in co-authorship networks based on hybrid content similarity metric," *Applied Intelligence,* vol. 48, no. 8, pp. 2470-2486, 2018.

[13]. E. Bütün, M. Kaya, and R. Alhajj, "Extension of neighbor-based link prediction methods for directed, weighted and temporal social networks," *Information Sciences,* vol. 463, pp. 152-165, 2018, https://doi.org/10.1016/j.ins.2018.06.051.

[14]. S. Behrouzi, Z. S. Sarmoor, K. Hajsadeghi, and K. Kavousi, "Predicting scientific research trends based on link prediction in keyword networks," *Journal of Informetrics,* vol. 14, no. 4, pp. 101079, 2020, https://doi.org/10.1016/j.joi.2020.101079.

[15]. A. Daud, W. Ahmed, T. Amjad, J. A. Nasir, N. R. Aljohani, R. A. Abbasi, and I. Ahmad, "Who will cite you back? Reciprocal link prediction in citation networks," *Library Hi Tech,* vol. 35, no. 4, pp. 509-520, 2017, https://doi.org/10.1108/LHT-02-2017-0044.

[16]. D. Liben-Nowell, and J. Kleinberg, "The link-prediction problem for social networks. journal of the Association for Information Science and Technology (2007)," *Google Scholar Google Scholar Digital Library Digital Library*, 2007,

[17]. S. Martinčić-Ipšić, E. Močibob, and M. Perc, "Link prediction on Twitter," *PloS one,* vol. 12, no. 7, pp. e0181079, 2017, https://doi.org/10.1371/journal.pone.0181079.

[18]. L. A. Adamic, and E. Adar, "Friends and neighbors on the web," *Social networks,* vol. 25, no. 3, pp. 211-230, 2003, https://doi.org/10.1016/S0378-8733(03)00009-1.

[19]. T. Zhou, L. Lü, and Y.-C. Zhang, “Predicting missing links via local information,” *The European Physical Journal B,* vol. 71, no. 4, pp. 623-630, 2009, https://doi.org/10.1140/epjb/e2009-00335-8.

[20]. M. E. Newman, “Clustering and preferential attachment in growing networks,” *Physical review E,* vol. 64, no. 2, pp. 025102, 2001, https://doi.org/10.1103/PhysRevE.64.025102

[21]. N. Benchettara, R. Kanawati, and C. Rouveirol, "A supervised machine learning link prediction approach for academic collaboration recommendation." pp. 253-256, https://doi.org/10.1145/1864708.1864760.

[22]. F. Almeida, and G. Xexéo, “Word embeddings: A survey,” *arXiv preprint arXiv:1901.09069*, 2019.

[23]. J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation." pp. 1532-1543.

[24]. J. Zhou, L. Liu, W. Wei, and J. Fan, “Network representation learning: from preprocessing, feature extraction to node embedding,” *ACM Computing Surveys (CSUR),* vol. 55, no. 2, pp. 1-35, 2022.

[25]. M. Grohe, "word2vec, node2vec, graph2vec, x2vec: Towards a theory of vector embeddings of structured data." pp. 1-16.

[26]. A. Grover, and J. Leskovec, "node2vec: Scalable feature learning for networks." pp. 855-864, https://doi.org/10.1145/2939672.2939754.

[27]. D. Lande, M. Fu, W. Guo, I. Balagura, I. Gorbov, and H. Yang, “Link prediction of scientific collaboration networks based on information retrieval,” *World Wide Web*, pp. 1-19, 2020, https://doi.org/10.1007/s11280-019-00768-9.

[28]. B. Liu, S. Xu, T. Li, J. Xiao, and X.-K. Xu, “Quantifying the Effects of Topology and Weight for Link Prediction in Weighted Complex Networks,” *Entropy,* vol. 20, no. 5, pp. 363, 2018, https://doi.org/10.3390/e20050363.

[29]. A. Hagberg, P. Swart, and D. S Chult, *Exploring network structure, dynamics, and function using NetworkX*, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

پیوست الف: الگوریتم node2vec

الگـوریتم node2vec بـر پایـه الگـوریتم word2vec کـه در پردازش متن برای تبدیل کلمات به فضای برداری استفاده می‌شـود ارائه شده است. در الگوریتم node2vec رأس‌ها معـادل کلمـات در الگوریتم word2vec هستند. الگوریتم word2vec دنبالـه‌ای از کلمات (جملات) را به عنوان ورودی می‌پذیرد. با انجـام تعـداد زیادی قدم زدن تصادفی و ثبت دنبالـه رأس‌هـای طـی شـده، ایـن دنباله برای رأس‌ها ساخته می‌شود (شکل الف-۱، قسمت ب).

در این الگوریتم نسخه خاصی از قدم زدن تصادفی به نام قدم زدن تصادفی درجه دوم[1] استفاده می‌شود. در قـدم زدن تصـادفی درجـه اول، تاریخچه قدم زدن نگهداری نمی‌شود. یعنی هنگامی که از یک رأس به رأس مبدأ انتقال صورت می‌گیرد، در مبدأ الگـوریتم نمـی داند از چه رأسی به حالت فعلی رسیده ولی در قدم زدن تصـادفی درجه دوم، رأسی که از آن به مبدأ انتقال صورت گرفته اهمیت پیدا می‌کند.

در قدم زدن تصادفی درجه دوم، احتمـال انتقـال از یـک رأس بـه رأس‌های مجاور از طریق تابع α که در معادله الـف-۱ نشـان داده شده است صورت می‌گیرد. در این رابطه t رأسی است که انتقـال از آن صورت گرفته، v رأسی است که در زمان فعلی الگـوریتم در آن قرار دارد و x رأس‌هایی که هستند که امکان انتقال به آن‌هـا از رأس v وجود دارد. احتمال بازگشت به رأس قبلـی p و احتمـال

[1] Second order random-walk

انتقال به قسمتی از گراف که تا به حال دیده نشده است q می باشد. همچنین d_{tx} کوتاه ترین فاصله بین t و x است.

$$\alpha_{pq}(t,x)=\begin{cases}\frac{1}{p} & if\ d_{tx}=0\\ 1\ if\ d_{tx}=1 & (p,q<1)\\ \frac{1}{q} & if\ d_{tx}=2\end{cases} \quad \text{معادله الف-۱}$$

با تغییر q و p می‌توان سرعت کاوش در گراف را تنظیم کرد. به عبارت دیگر، با کاهش q قدم زدن تصادفی به سمت پیمایش عمقی[1] گراف متمایل می‌شود و با کاهش p الگوریتم به سمت پیمایش سطحی[2] گراف متمایل می‌شود.

در ادامه هر یک از این دنباله‌های ساخته شده توسط قدم زدن تصادفی به عنوان ورودی به شبکه عصبی skip-gram داده می شود. مدل skip-gram یک شبکه عصبی با یک لایه پنهان است. این مدل آموزش داده می‌شود که احتمال وجود یک رأس (کلمه) را در صورت وجود یک رأس دیگر در دنباله محاسبه کند. خروجی مدل skip-gram یک نهفته سازی برای هر رأس است (شکل الف-۱، قسمت ج و د).

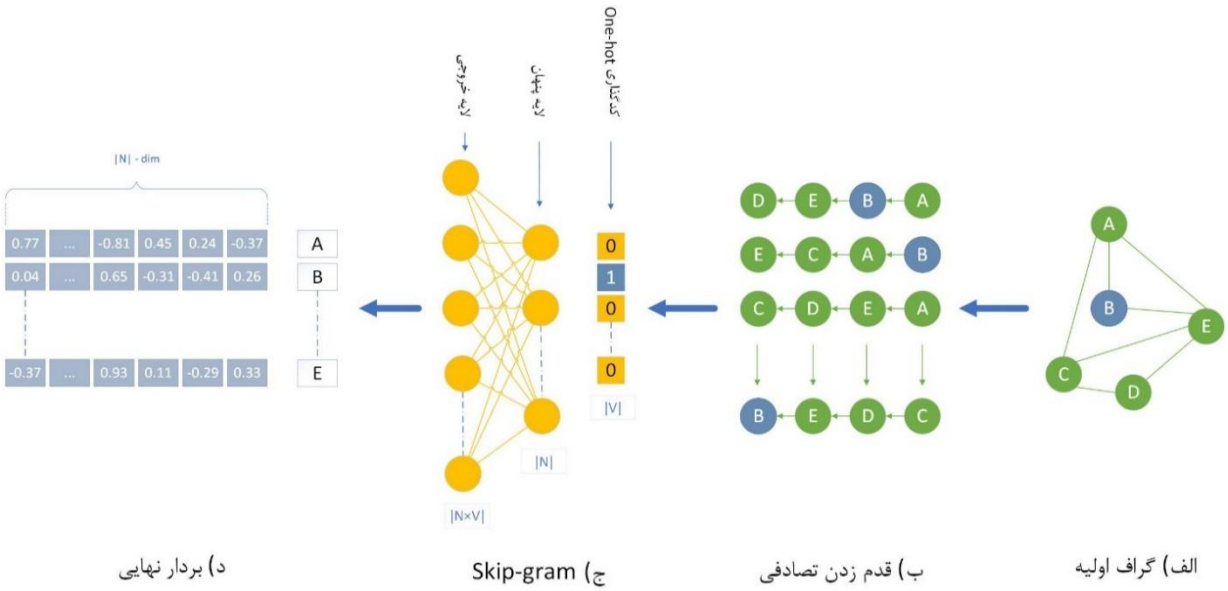
پیوست ب: چارچوب OGB

OGB یک چارچوب نرم‌افزاری متن باز برای انجام آزمایشات و ارزیابی نتایج برای مسائل مربوط به پردازش گراف است. این چارچوب با زبان پایتون و توسط دانشگاه استنفورد توسعه داده شده‌است[۲۷]. به طور کلی این چارچوب ارائه مجموعه داده‌ها و ارزیابی نتایج را در یک قالب استاندارد برای پژوهشگران فراهم می‌کند. در نهایت بهترین نتایج ثبت شده که موفق به ثبت مقاله شده باشند در جدول برترین نتایج در سایت OGB نمایش داده می‌شود. مراحل انجام یک پژوهش در چارچوب OGB در شکل‌ب-۱ نشان داده شده است. ابتدا یک مجموعه داده برای پژوهش مورد نظر انتخاب می‌شود. سپس OGB این مجموعه داده را به سه قسمت آموزشی، اعتبارسنجی و آزمایشی تقسیم و بارگزاری می‌کند. سپس با استفاده از یک الگوریتم داده‌های اعتبارسنجی و آزمایشی پردازش می‌شوند و در نهایت نتایج توسط ارزیاب مختص به مجموعه داده انتخاب شده ارزیابی می‌شود و در صورت انتشار مقاله در تابلوی امتیازات درج می‌شود.

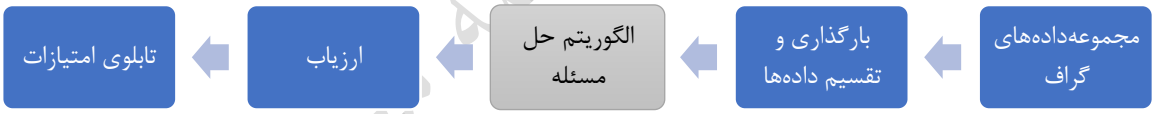
مسائلی که در OGB برای آن‌ها مجموعه داده و ارزیاب فراهم شده اعم از پیش‌بینی پیوند، دسته‌بندی رأس ، تکمیل گراف دانش و مسائل دیگر هستند. این مجموعه‌داده‌ها در زمینه‌هایی مانند شبکه‌های علمی، گراف‌های بیولوژیکی و شبکه‌های دانش هستند. در این بخش فقط به مجموعه‌داده‌های مرتبط با شبکه‌های علمی و مسئله پیش‌بینی پیوند می‌پردازیم.

استفاده از چنین چارچوب‌هایی به پژوهشگران کمک می‌کند نتایج پژوهش‌های خود را با سایر الگوریتم‌های پیشین مقایسه کنند. در این مقاله آزمایشات و ارزیابی نتایج مبتنی بر این چارچوب انجام می‌شود. به عنوان نمونه تابلوی نمایش و رتبه‌بندی ارزیابی تعدادی از مقالات، بر اساس معیار Hits@50 با مجموعه داده ogbl-collab در شکل ب-۲ نشان داده شده است.

[1] DFS
[2] BFS



شکل الف-۱) الگوریتم node2vec



شکل ب-۱) مراحل انجام یک مسئله گراف در OGB

Rank	Method	Ext. data	Test Hits@50	Validation Hits@50	Contact	References	#Params	Hardware	Date
1	**GIDN@YITU**	No	0.7096 ± 0.0055	0.9620 ± 0.0040	Zixiao Wang, Yu Zhang(ZhejiangLab, HUST)	Paper, Code	60,449,025	DepGraph@SCTS/CGCL	Oct 10, 2022
2	**PLNLP + SIGN**	No	0.7087 ± 0.0033	1.0000 ± 0.0000	Liang Yao (Tencent)	Paper, Code	34,980,864	Tesla-P40 (24G GPU)	Apr 7, 2022
3	**PLNLP (random walk aug.)**	No	0.7059 ± 0.0029	1.0000 ± 0.0000	Zhitao Wang (WeChat@Tencent)	Paper, Code	34,980,864	Tesla-P40 (24G GPU)	Dec 21, 2021
4	**HOP-REC**	No	0.7012 ± 0.0016	1.0000 ± 0.0000	Bo-Yu Lin (CFDA & CLIP Labs)	Paper, Code	30,191,104	CPU	Oct 21, 2021
5	PLNLP+ LRGA	No	0.6909 ± 0.0055	1.0000 ± 0.0000	Hao Xu	Paper, Code	35,200,656	NVIDIA Tesla V100 (32GB GPU)	Jun 26, 2022
17	DeepWalk	No	0.5037 ± 0.0034	Please tell us	Hao Xiong (DGL)	Paper, Code	61,390,187	g4dn.2xlarge, T4 (15GB GPU)	Jun 30, 2020
18	Node2vec	No	0.4888 ± 0.0054	0.5703 ± 0.0052	Matthias Fey – OGB team	Paper, Code	30,322,945	GeForce RTX 2080 (11GB GPU)	Jun 22, 2020
19	GraphSAGE	No	0.4810 ± 0.0081	0.5688 ± 0.0077	Matthias Fey – OGB team	Paper, Code	460,289	GeForce RTX 2080 (11GB GPU)	Jun 24, 2020
20	GCN (val as input)	No	0.4714 ± 0.0145	0.5263 ± 0.0115	Matthias Fey – OGB team	Paper, Code	296,449	GeForce RTX 2080 (11GB GPU)	Oct 19, 2020
21	GCN	No	0.4475 ± 0.0107	0.5263 ± 0.0115	Matthias Fey – OGB team	Paper, Code	296,449	GeForce RTX 2080 (11GB GPU)	Jun 24, 2020
22	Matrix Factorization	No	0.3886 ± 0.0029	0.4896 ± 0.0029	Matthias Fey – OGB team	Paper, Code	60,514,049	GeForce RTX 2080 (11GB GPU)	Jun 22, 2020

شکل ب-۲: تابلوی امتیازات برای مجموعه‌داده **ogbl-collab**