Journal of Soft Computing and Information Technology (JSCIT)
Babol Noshirvani University of Technology, Babol, Iran
Journal Homepage: jscit.nit.ac.ir
Volume 11, Number 2, Summer 2022, pp. 41-52
Received: 12/08/2021, Revised: 05/16/2022, Accepted: 06/07/2022

New Criteria for Predicting Links based on Node Composition and Network Structure

Hasan Saeidinezhad [1], Elham Parvinnia [*2]
1- Department of Computer Engineering, Shiraz Branch, Islamic Azad University, Shiraz, Iran
2- Department of Computer Engineering, Shiraz Branch, Islamic Azad University, Shiraz, Iran
[1]saeidinezhade@gmail.com, [2*]parvinnia@iaushiraz.ac.ir

Corresponding author address: Elham Parvinnia, Faculty of Computer Engineering, Shiraz Branch, Islamic Azad University, Shiraz, Iran

Abstract - Currently, the study of social space and social networks and the analysis of these networks has grown significantly. In real life, people are not independent of each other. People in social groups are interdependent. One of the most widely used fields in the study of social networks is the issue of link identification, which has recently become popular among domestic and foreign researchers. Link prediction can be used not only in the field of social networking, but also in areas such as bioinformatics, to explore the interrelationships between proteins, in the field of e-commerce. The main purpose in the field of link identification is to investigate the possibility of creating or deleting links between members in the future state of the network using the analysis of its current state. In this study, using local criteria of neighbor-based similarity and general path-based similarity criteria, both of which use the graph structure, new similarity criteria have been introduced. The results of the work have been tested based on precision and AUC criteria on the data set and show the superiority of the proposed method that uses a combination of graph structure information like path, neighbor and node degrees over the criteria that use only the path or the neighbor.

Keywords - social networks, link identification, network structure, path similarity criterion, community

مجله علمی پژوهشی رایانش نرم و فناوری اطلاعات
دانشگاه صنعتی نوشیروانی بابل
صفحه مجله: jscit.nit.ac.ir
جلد ۱۱، شماره ۲، تابستان ۱۴۰۱ ، صفحه ۴۱-۵۲
دریافت: ۱۴۰۰/۰۹/۱۷، بازنگری: ۱۴۰۱/۰۲/۲۶، پذیرش: ۱۴۰۱/۰۳/۱۷

معیارهای جدید جهت پیش‌بینی لینک مبتنی بر ترکیب نودها و ساختار شبکه

حسن سعیدی نژاد[۱]، الهام پروین نیا*[۲]
۱- دانشکده مهندسی کامپیوتر ـ دانشگاه آزاد اسلامی شیراز، شیراز، ایران
*۲- دانشکده مهندسی کامپیوتر ـ دانشگاه آزاد اسلامی شیراز، شیراز، ایران
[1]saeidinezhade@gmail.com, [2*]parvinnia@iaushiraz.ac.ir

* نشانی نویسنده مسئول: الهام پروین نیا، شیراز، کیلومتر ۵ جاده صدرا، پردیس دانشگاه آزاد اسلامی واحد شیراز، دانشکده کامپیوتر.

چکیده- امروزه مطالعه بر روی فضای اجتماعی و شبکه‌های اجتماعی و تحلیل و بررسی این شبکه‌ها رشد چشم‌گیری داشته است. در زندگی واقعی، افراد مستقل از یکدیگر نیستند. افراد در گروه‌های اجتماعی به هم وابسته شده‌اند. از جمله پرکاربردترین زمینه‌ها در مطالعه‌ی شبکه‌های اجتماعی، بحث شناسایی لینک[۱] می‌باشد که اخیراً بسیار مورد استقبال پژوهشگران داخلی و خارجی قرار گرفته است. پیش‌بینی لینک نه تنها می‌تواند در زمینه شبکه اجتماعی استفاده شود بلکه در زمینه‌هایی چون بیوانفورماتیک، برای کشف روابط متقابل بین پروتئین‌ها مورد استفاده قرار می‌گیرد. هدف اصلی شناسایی لینک، بررسی احتمال ایجاد یا حذف لینک بین اعضا در وضعیت آینده‌ی شبکه با استفاده از تحلیل وضعیت کنونی آن است. در این پژوهش با بهره‌گیری از معیارهای محلی شباهت مبتنی بر همسایه[۲] و معیار عمومی شباهت مبتنی بر مسیر[۳] که هر دو از ساختار گراف[۴] استفاده می‌کنند معیارهای شباهت جدیدی معرفی شده است. نتایج کار بر روی مجموعه داده‌های مورد بررسی برتری کار را نسبت به معیارهایی که تنها از مسیر و یا از همسایه بهره می‌جویند، نشان می‌دهد و مشاهده می شود که متد ارائه شده با ترکیب اطلاعات مسیر و همسایه‌ها می‌تواند بر اساس معیارهای precision و AUC با دقت بیشتری نسبت به روش‌های قبل لینک‌ها را پیش‌بینی کند.

واژه‌های کلیدی: شبکه‌های اجتماعی، شناسایی لینک، ساختار شبکه، معیار شباهت مسیر، جامعه

۱- مقدمه

شبکه‌ها[۵]، یک روش برای تحلیل و بررسی ساختارهای اجتماعی هستند که تمام توجه خود را بر روی روابط بین اعضایی که این‌گونه ساختار اجتماعی را شکل می‌دهند جلب کرده‌اند. به این اعضا نود[۶] یا راس گفته می‌شود که این نودها دارای ویژگی‌هایی هستند که آن‌ها را از هم متمایز می‌کند [۱]. به‌عنوان مثال در شبکه‌ی افراد یک سازمان، ویژگی‌های نود را می‌توان مذکر یا مؤنث بودن و یا سمت هر شخص در آن سازمان برشمرد؛ بنابراین وقتی از شبکه‌های اجتماعی[۷] صحبت به میان آورده می‌شود منظور بستری اجتماعی[۸] است که شامل رئوسی (فردی یا سازمانی) است که توسط یک یا چند نوع خاص از رابطه مانند دوستی، خویشاوندی، سرایت بیماری و ... به هم وصل هستند که به این روابط یال[۹] گفته می‌شود. در ساده‌ترین شکل، یک شبکه‌ی اجتماعی نگاشتی است که رئوس را به‌وسیله‌ی یال‌های مربوط به هم متصل می‌کند؛ رئوس، بازیگران[۱۰] درون شبکه و یال‌ها روابط میان این بازیگران است [۲].

شبکه‌های اجتماعی در زمینه‌های متفاوتی مورد مطالعه و بررسی قرار گرفته است. از جمله پرکاربردترین حوزه‌های مطالعاتی در شبکه‌های اجتماعی، شناسایی لینک[۱۱] است. مسئله‌ای که حوزه‌ی شناسایی لینک درصدد حل آن است، این است که: با داشتن یک اسنپ شات از وضعیت فعلی شبکه آیا می‌توان پیش‌بینی کرد که بین کدامیک از اعضای شبکه در اسنپ شات بعدی از شبکه ممکن است رابطه‌ای به وجود بیاید؟ [۳]

اهمیت این مسئله در سیستم‌های پیشنهادی[۱۲] یا فروش اینترنتی[13] آشکار می‌شود که به افراد امکان پیدا کردن محصولات مورد علاقه را می‌دهد [۴]، یا به آن‌ها کمک می‌کند تا دوستانی جدید پیدا کنند [۵]، در شبکه‌های اجتماعی آکادمیک که به افراد امکان پیدا کردن نویسنده همکار یا متخصص می‌دهد [۶]، یا در مقیاس بزرگی از شبکه‌های ارتباطی می‌توان مخاطبین یک فرد خاص را بر روی تلفن همراه پیش‌بینی کرد [۷]. همچنین می‌توان از این حوزه‌ی تحقیقاتی برای کامل کردن یک شبکه با بهره‌گیری از اطلاعات ناقص و جزئی آن شبکه استفاده کرد [۹ و ۸] و مسیر تکامل یک شبکه را بهتر فهمید [۱۰]. علاوه بر این حوزه‌ی شناسایی لینک، در علوم مربوط به بیوانفورماتیک[۱۴] و زیست‌شناسی[۱۵] نیز بسیار پرکاربرد است، به‌عنوان مثال در پیش‌بینی خصوصیاتی که احتمال راه یافتن آن‌ها به آینده بیشتر است، یا شناسایی فعل‌وانفعالات بین پروتئین‌ها و همچنین در شبکه‌های بیان ژن[۱۶] [۱۱] می‌توان از حوزه‌ی شناسایی لینک بهره جست.

تکنیک‌های موجود در حوزه‌ی شناسایی لینک به ۳ دسته‌ی مبتنی بر نود[۱۷]، مبتنی بر ساختار[۱۸] و مبتنی بر تئوری اجتماعی[۱۹] تقسیم می‌شوند [۱۲]. تکنیک‌های node-based بر اساس ویژگی‌های منحصربه‌فرد هر یک از نودهای شبکه شباهت بین نودها را تخمین می‌زنند، به عنوان مثال در شبکه‌ی فروش اینترنتی، ویژگی‌هایی مانند جنسیت، سن و کالاهای خریداری کرده و ... می‌تواند برای هر فرد (نود) در نظر گرفته شود و بر اساس این ویژگی‌ها افراد مشابه به هم را پیدا کرد و به افراد کالای جدیدی که مطابق میل آن‌ها باشد را پیشنهاد داد. تکنیک‌های مبتنی بر ساختار[۲۰] از ساختار گراف بهره می‌جویند و خود به ۳ دسته‌ی مبتنی بر همسایه[۲۱]، مبتنی بر مسیر[۲۲] و مبتنی بر قدم تصادفی[۲۳] تقسیم می‌شوند [۱۲]. از مهم‌ترین الگوریتم‌های مبتنی بر همسایه مانند Common neighbor (CN)، preferential attachment (PA) و Adamic Adar (AA) هستند که از همسایگان مستقیم نودها استفاده می‌کنند تا بر اساس آن تخمینی از میزان شباهت ۲ نود به دست بیاورند.

برای مثال الگوریتم Common Neighbor بر این اساس که هرچه میزان همسایگان مشترک بین ۲ نود بیشتر باشد احتمال ایجاد لینک بین آن ۲ نود بیشتر می‌شود، معیاری از شباهت بین دو نود x و y را به‌صورت زیر ارائه داده است [۲۵]:

$$CN(x,y) = |\Gamma(x) \cap \Gamma(y)| \qquad (۱)$$

که $\Gamma(x)$ مجموعه‌ی همسایگان نود x می‌باشد و $|\Gamma(x) \cap \Gamma(y)|$ بیانگر تعداد همسایگان مشترک نود x و y است.

الگوریتم Preferential Attachment حاصل‌ضرب درجه‌های دو نود موردنظر را به عنوان معیاری از شباهت آن دو نود در نظر می‌گیرد، با این فرض که هرچه درجه‌ی ۲ نود بیشتر باشد احتمال ایجاد لینک بین آن دو نود نیز افزایش می‌یابد. معیاری از شباهت بین دو نود x و y را به‌صورت زیر ارائه داده است [۱۰]:

$$PA(x,y) = |\text{Degree}(x)| \,.\, |\text{Degree}(y)| \qquad (۲)$$

که $|\text{Degree}(x)|$ و $|\text{Degree}(y)|$ به ترتیب درجه‌ی نود x و y می‌باشد.

همچنین الگوریتم Adamic Adar که در بسیاری از کارهای صورت گرفته مورد استفاده قرار گرفته است، بر این اساس است که هرچه تعداد همسایگان مشترک ۲ نود x و y، تعداد همسایگان کمتری داشته باشند آنگاه احتمال ایجاد لینک بین x و y افزایش می‌یابد و به این صورت محاسبه می‌شود [۱۳]:

$$AA(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log|\Gamma(z)|} \qquad (۳)$$

که z نودهای موجود در مجموعه‌ی همسایگان مشترک x و y است و $|\Gamma(z)|$ درجه‌ی نود z است.

از جمله مهم‌ترین الگوریتم‌های مبتنی بر مسیر می‌توان به الگوریتم Katz اشاره کرد. در محاسبه‌ی الگوریتم katz تمامی مسیرهای موجود بین دو نود x و y محاسبه شده و در الگوریتم دخالت داده می‌شود. احتمال ایجاد لینک بین x و y افزایش می‌یابد و به این صورت محاسبه می‌شود:

$$Katz(x,y) = \sum_{l=1}^{\infty} \beta^l . |path_{x,y}^l| = \beta A + \beta^2 A^2 + \cdots \qquad (۴)$$

$\beta > 0$ پارامتری است که میزان تأثیرگذاری مسیرهای به طول زیاد را تعیین می‌کند. هرچه β مقدار کوچک‌تری داشته باشد الگوریتم Katz عملکردی مشابه به Common Neighbor از خود نشان می‌دهد زیرا تأثیر مسیرهای به طول زیاد در محاسبه‌ی شباهت بسیار کم می‌شود [۱۴].

الگوریتم‌هایی که ذکر شد الگوریتم‌هایی هستند که عملکرد آن‌ها در این مقاله در قسمت نتایج مورد بررسی قرار گرفته همچنین با

معیار شباهت CNDP ترکیب شده‌اند و نتایج با هم مقایسه شده است.

شباهت بین دو نود توسط معیار CNDP به‌صورت زیر محاسبه می‌شود [۱۵]:

$$CNDP\ (x,y) = \sum_{z\ \epsilon \Gamma(x) \cap\ \Gamma(y)} |C_{(z)}|.|\Gamma(z)|^{-\beta C} \qquad (۵)$$

که در این رابطه z نودی است در مجموعه‌ی همسایگان مشترک x و y، $|C_{(z)}|$ تعداد همسایگان مشترک z و x و y، $|\Gamma(z)|$ درجه‌ی نود z ، C میانگین clustering coefficient در گراف و β یک مقدار ثابت است که با آزمایش روی دیتاست‌های مختلف مقدار بهینه‌ی آن از روش درون‌یابی[۲۴] به دست آمده است.

در این مقاله، برای محاسبه‌ی احتمال ایجاد لینک بین دو نود x و y، ابتدا از معیار شباهت CNDP کمک می‌گیریم که علاوه بر همسایگان مستقیم از همسایگان (درجه ۲) نیز برای محاسبه‌ی شباهت استفاده می‌کند و با استفاده از این معیار اطلاعات همسایگان بین دو نود x و y را به دست می‌آوریم. سپس با بهره‌گیری از معیارهایی مانند PA و Katz اطلاعات مسیر بین دو نود x و y را به دست می‌آوریم . با ترکیب اطلاعات مبتنی بر مسیر و اطلاعات مبتنی بر همسایه، معیاری از شباهت به دست می‌آید که با تنظیم پارامترهای میزان تأثیرگذاری هر معیار به دقتی بهتر از دقت معیارهای قبل که تنها از یک منبع اطلاعات بهره می‌گرفتند می‌رسیم. در واقع روش پیشنهادی این مقاله ترکیب روش‌های موجود قبلی است که با بهره‌گیری از چند منبع اطلاعات ساختاری گراف، دقت بهتری در پیش‌بینی لینک حاصل شود. نتایج گزارش شده در بخش ۴ مقاله، برتری الگوریتم ارائه شده را نسبت به الگوریتم CNDP و سایر الگوریتم‌های متناظر نشان می‌دهد.

ادامه‌ی مقاله به این صورت است: در بخش ۲ کارهای پیشین مرور شده است، در بخش ۳ الگوریتم ارائه شده به‌تفصیل توضیح داده خواهد شد در بخش ۴ نتایج، بررسی و تحلیل خواهد شد و در بخش آخر از مقاله نتیجه‌گیری و کارهای آتی بیان خواهد شد.

۲- کارهای پیشین

از جمله الگوریتم‌های مطرح در این حوزه ADP[۲۵] است [۱۶]. در این الگوریتم درجه‌ی همسایگان مشترک بر اساس مقدار clustering coefficientی که شبکه‌ی مورد بررسی دارد، محدود[۲۶] می‌شود. هدف از این کار این است که الگوریتم شباهتی که ارائه می‌شود، بتواند برای تمام شبکه‌ها با ساختارهای متفاوت، دقت قابل قبولی در پیش‌بینی لینک داشته باشد. بر این اساس در روش ADP به‌منظور آنکه از ساختار منحصربه‌فرد شبکه‌ی مورد بررسی نیز در محاسبه‌ی شباهت استفاده شود الگوریتم زیر ارائه داده شده که پارامتر c بیانگر مقدار ACC شبکه‌ی مورد نظر می‌باشد و پارامتر β مقدار ثابت است که مقدار ۲.۵ برای آن در نظر گرفته شده است.

$$S\ (x,y) = \sum_{z\ \epsilon \Gamma(x) \cap\ \Gamma(y)} |\Gamma(z)|^{-\beta C} \qquad (۶)$$

الگوریتم دیگری به نام NCCCN در [۱۷] مطرح شد که در این مقاله اطلاعات شبکه‌ی خوشه‌بندی شده با اطلاعات همسایگان مشترک بین دو نود ترکیب شد و معیاری از شباهت بین ۲ نود، بر اساس این دو ۲ منبع اطلاعات به دست آمد. همچنین در [۱۸]معیار شباهتی مبتنی بر همسایگان به نام triadic measure معرفی شد که از واحدهایی به نام موتیف[۲۷] استفاده می‌کند. موتیف‌ها فرم‌های گوناگونی از شبکه‌های کوچک‌اند که شامل ۳ نود در یک شبکه‌ی جهت‌دار هستند. در این مقاله ۱۳ نوع موتیف مختلف معرفی شده است. الگوریتم معرفی شده در این مقاله برای محاسبه‌ی شباهت بین ۲ نود x و y به این صورت است که به ازای هر یک از همسایگان مشترک بین دو نود مانند z، تعداد موتیف-های متشکل از ۳ نود x و y و z شمرده می‌شود و حاصل بر ۱۳ تقسیم می‌شود و این روند برای تمام همسایگان مشترک محاسبه شده و در نهایت با هم جمع می‌شود، سپس عدد حاصل بر تعداد کل همسایگان مشترک بین دو نود x و y تقسیم می‌شود. نتیجه‌ی به دست آمده نشان دهنده‌ی میزان شباهت بین ۲ نود x و y است.

اطلاعات خوشه‌بندی شبکه ، اطلاعات بسیار مفیدی در روند شناسایی لینک است [۲۲] . به همین منظور نویسندگان در [۲۳] شبکه‌ی مورد بررسی را خوشه‌بندی کردند و مشاهده کردند که رابطه‌ی زیادی بین این خوشه‌ها که برگرفته از ساختار گراف است و دقت الگوریتم در شناسایی لینک وجود دارد.

همچنین در [۲۴] بسطامی[۲۸] و همکاران روشی برای شناسایی لینک مبتنی بر جاذبه[۲۹] ارائه دادند که از اطلاعات خوشه‌ها در شبکه نیز استفاده می‌کرد. آن‌ها به‌صورت موازی بر روی خوشه‌های شناسایی شده الگوریتمشان را اعمال کردند تا سرعت اجرای کار افزایش یابد. همچنین از معیار شباهت Adamic Adar برای محاسبه‌ی شباهت بین نودها استفاده کردند.

بسیاری از مطالعات صورت گرفته بر روی ترکیب ویژگی‌های ساختاری شبکه کار کرده‌اند و مشاهده شده است که ویژگی‌های مبتنی بر نود می‌تواند در محاسبه‌ی شباهت بسیار مفید باشد [۱۹]. در [۲۰] ین و همکاران[۳۰] یک فریم ورک طراحی کردند که هم از ساختار گراف و هم از ویژگی‌های منحصربه‌فرد نودها برای شناسایی لینک استفاده می‌کرد. آن‌ها کار خود را بر روی دیتاست‌های Co-authorship و co-starring مورد ارزیابی قرار دادند و مشاهده کردند که از ترکیب اطلاعات گراف نتایج بسیار بهتری حاصل می‌شود به نسبت وقتی که از یک ویژگی به‌تنهایی استفاده گردد. آن‌ها از پارامتر γ برای تنظیم میزان دخالت هر یک از ویژگی‌ها، استفاده کردند.

در مطالعه‌ی دیگری [۲۱] ونگ و همکاران[۳۱] اذعان کردند که شباهت بین فعالیت‌های دو نود با فاصله‌ی بین دو نود متناسب است، به این معنی که هرچه فاصله‌ی بین دو نود کمتر باشد میزان شباهت آن‌ها بیشتر خواهد بود. آن‌ها ویژگی‌های منحصربه‌فرد نود مانند فعالیت[۳۲] و مسیر حرکت آن را با ویژگی‌های ساختاری شبکه ترکیب کردند و یکsupervised classifier طراحی کردند که دقت در میزان پیش‌بینی را به‌طور محسوسی افزایش می‌داد.

در [۲۶] یک دیتاست واقعی مبتنی بر شبکه‌ی ارتباطات تلفنی ارائه شد که ساختاری درختی دارد و بسیار اسپارس می‌باشد. در این مقاله تعدادی از روش‌های پیش‌بینی لینک مانند AdamicAdar، Common Neighbors، Node2Vec، NeoGNN و SEAL بررسی شد. مدل‌های مبتنی بر هیوریستیک مانند AA و CN برای این دیتاست به دلیل اسپارسیتی بالا، ناکارآمد بود. اما مدل‌های NeoGNN و SEAL به دقت خوبی رسیدند.

در [۲۷] فریم ورکی بر پایه‌ی بر شبکه‌های عصبی مبتنی بر گراف ارائه شد که به‌منظور پیش‌بینی لینک داده‌های بیولوژیکی مانند رشته‌های پروتئینی و ساختار مولکول‌های دارویی مورد استفاده قرار گرفت . در مدلی که ارائه دادند از شبکه‌های عصبی CNN و GCN برای یادگیری ویژگی‌های نودهایی که از دیتاست‌های بیولوژیکی استخراج می‌شود استفاده شده است.

در [۲۸] روشی نوین برای شناسایی لینک بر پایه معیارهای متفاوت مرکزیت نود ارائه شد و از کلاسیفایرهای مختلف به‌منظور تشخیص ایجاد یا عدم ایجاد لینک استفاده شد. با استفاده از معیارهای مختلف مرکزیت نود می‌توان ساختارهای محلی، نیمه محلی و کلی شبکه‌ها را مشخص کرد. همچنین از این معیارهای مرکزیت نود به عنوان ویژگی‌های نود استفاده کردند. وجود یا عدم وجود لینک نیز به معنای برچسب مثبت یا منفی برای هر نود محسوب می‌شود. بنابراین دیتاستی برچسب دار فراهم شد که به‌وسیله‌ی کلاسیفایر light gradient boosted machine توانستند مدلی طراحی کنند و با آن لینک‌ها را پیش‌بینی کنند.

جدول ۱: مقایسه مقالات مرتبط با حوزه‌ی شناسایی لینک

مقاله	نوآوری مقاله	معیارهای ارزیابی
[15]	ارائه‌ی معیاری از شباهت بر اساس همسایگان مشترک از درجه اول و دوم و CO اختصاصی هر شبکه	ADP, CNDP, NCCCN, Triadic measure
[16]	از ویژگی‌های منحصربه‌فرد شبکه مانند CO استفاده کردند تا برای هر شبکه بر اساس ساختار دقت قابل قبولی کسب شود	ADP, CN, AA, RA
[17]	ارائه معیاری ترکیبی از اطلاعات همسایگان مشترک و اطلاعات خوشه‌های شبکه	NCCCN, ENCCCN, ENCCPT, NCCPT, CLAPA, CN, AA, PA
[18]	با ترکیب‌های ۳ تایی مختلف نودهای جهت‌دار در شبکه معیاری به نام triadic measure ارائه دادند.	Triadic measure, AA, PA, CN, JC
[20]	معیاری از شباهت با ترکیب ویژگی‌های ساختاری گراف و ویژگی نودها	SimAttr, Katz SVM, LINKREC, AA
[21]	ارائه روشی با نظارت بر اساس ویژگی‌های نود و فاصله دو نود	Co-Location rate(CoL), Weited Co-Location rate (WCoL)
[23]	استفاده از اطلاعات خوشه‌ها در شبکه‌های پیچیده برای شناسایی لینک	CN, AA, RA, SRW, Katz, PR
[24]	یک روش پیش‌بینی لینک مبتنی بر گرانش در شبکه‌های اجتماعی	AA, Jaccard, CN
[26]	ارائه‌ی دیتاستی اسپارس مبتنی بر شبکه‌ی ارتباطات و بررسی الگوریتم‌های مختلف LP روی آن	Adamic Adar, CN, node 2 vec, NeoGNN, SEAL
[27]	ارائه‌ی یک فریمورک بر مبنای شبکه‌های عصبی برای تشخیص پیوند بین داده‌های بیولوژکی	AA, CN, PA, Jaccard
[28]	استفاده از کلاسیفایرها برای پیش‌بینی لینک در شبکه، ویژگی‌های نود به عنوان attribute و وجود یا عدم وجود لینک به عنوان لیبل	Adamic Adar, CN, PA
Proposed method	ارائه‌ی معیاری ترکیبی بر اساس اطلاعات همسایگان مشترک و مسیر و درجه‌ی نودها	PNS, DPNS, DNS, ADP, NCCCN, CNDP, Triadic measure

۳- روش ارائه داده شده

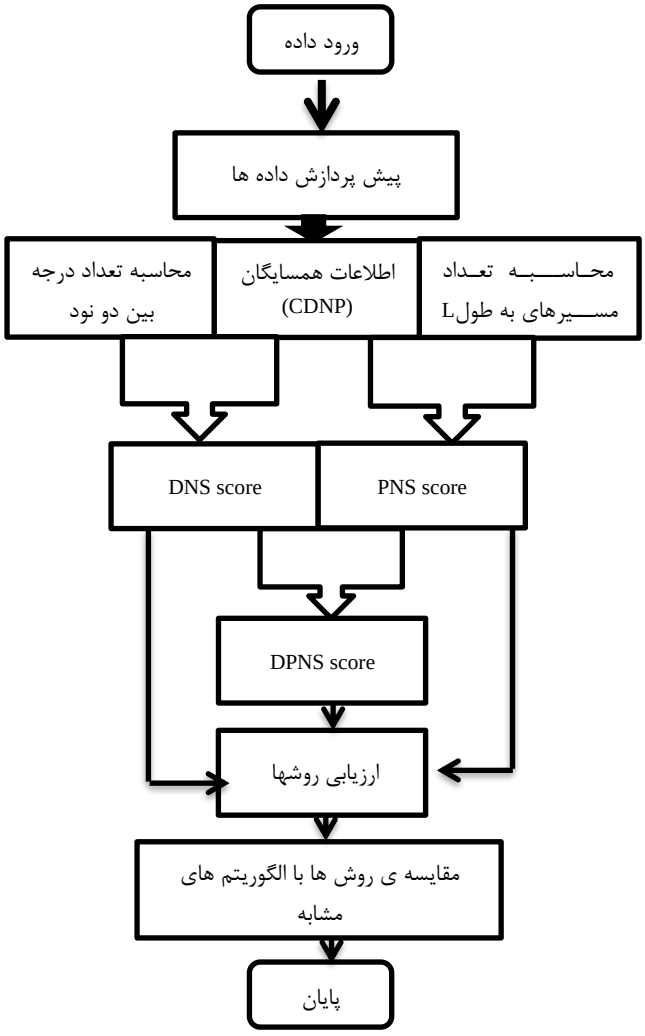
فریمورک اصلی تمام الگوریتم‌های پیش‌بینی لینک مبتنی بر شباهت، مشابه هم است و تنها تفاوت در نحوه‌ی محاسبه‌ی معیار شباهت می‌باشد. به همین دلیل تمامی الگوریتم‌های ارائه شده سعی بر این داشته‌اند تا با تغییراتی در نحوه‌ی محاسبه‌ی معیار شباهت بتوانند بهبودی بر روی دقت و صحت پیش‌بینی داشته باشند [۵]. بر این اساس در این مقاله معیاری از شباهت ارائه شده که به نسبت دیگر معیارهای مبتنی بر شباهت مانند ADP، NCCCC و Triadic measure و همچنین CNDP، بر روی دیتاست‌های مشترک بهتر عمل می‌کند. در این معیار سعی شده تا علاوه بر بهره‌گیری از اطلاعات همسایگان مشترک دو نود، از مسیر بین دو نود و همچنین از درجه‌ی دو نود مبدأ و مقصد نیز استفاده شود تا با ترکیب این اطلاعات ساختاری بتوان دقیق‌تر از روش‌های قبل که تنها از یک ویژگی گراف استفاده می‌کردند، لینک‌ها را پیش‌بینی کرد.

فلوچارت روش پی شنهادی در شکل ۱ آمده ا ست بدین منظور، ابتدا شبکه‌ی مورد بررسی، به‌صورت گراف G = (V,E) پیاده‌سازی شده که V مجموعه‌ی رئوس و E مجموعه‌ی یال‌های بین رئوس می‌با شد. سپس گراف G مورد پیش‌پردازش[۳۳] قرار می‌گیرد تا در صورتی که یال‌ها جهت‌دار[۳۴] با شند آن‌ها را به یال‌های بدون جهت[۳۵] تبدیل کند و همچنین اگر گراف به صورت multi edge ا ست طی پیش‌پردازش به گراف ساده[۳۶] تبدیل می شود. و سپس بر روی گراف ساده و بدون جهت G الگوریتم شنا سایی لینک به‌صورت زیر اعمال می‌شود.

برای محاسبه‌ی شباهت بین دو نود بر اساس ترکیب اطلاعات مسیر و همسایگان (PNS)، طبق معیار شباهت زیر عمل می‌کنیم:

$$path\_neighbor\_Similarity\ (x,y) = \sum_{l=3}^{5} \alpha^{l}.|path_{x,y}^{l}| \times \sum_{z\ \epsilon\Gamma(x)\cap\Gamma(y)} \frac{|C_{(z)}|}{|\Gamma(z)|^{\beta C}} \quad (۷)$$

که در این رابطه α ضریبی است که تأثیر مسیرهای به طول زیاد را تعیین می‌کند، مقادیر کوچک α تأثیر مسیرهای طولانی را کم می‌کند. $|path_{x,y}^{l}|$ بیانگر تعداد مسیرهای به طول l بین دو نود x و y است. به دلیل پیچیدگی محاسباتی[۳۷] در پیدا کردن مسیرهای به طول زیاد، تنها مسیرهای به طول بین ۳ تا ۵ بررسی شده است. همچنین $|C_{(z)}|$ تعداد همسایگان مشترک z و x و y، $|\Gamma(z)|$درجه‌ی نود z ، C میانگین clustering coefficient در گراف و β یک مقدار ثابت است که با آزمایش روی دیتاست‌های مختلف مقدار بهینه‌ی آن از روش درون‌یابی[۳۸] به دست آمده است.



شکل ۱: فلوچارت روش پیشنهادی

همچنین معیار دیگری بر پایه‌ی درجه‌ی نود و همسایگان (DNS) ارائه شده که به‌صورت زیر است:

$$\text{degree\_neghibor Similarity}\ (x,y) = x\_degree \times y\_degree \times \sum_{z\ \epsilon\Gamma(x)\cap\Gamma(y)} \frac{|C_{(z)}|}{|\Gamma(z)|^{\beta C}} \quad (۸)$$

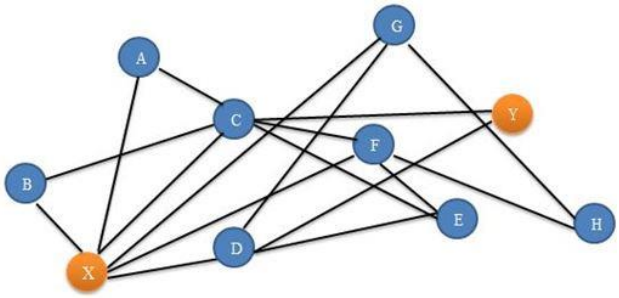
x_degree درجه‌ی نود x ، y_degree درجه‌ی نود y است. $|C_{(z)}|$ تعداد همسایگان مشترک z و x و y، $|\Gamma(z)|$ درجه‌ی نود z ، C میانگین clustering coefficient در گراف و β یک مقدار

ثابت است که با آزمایش روی دیتاست‌های مختلف مقدار بهینه‌ی آن از روش درون‌یابی به دست آمده است. در این معیار، از حاصل‌ضرب درجه‌ی دو نود مورد نظر در تعداد همسایگان مشترک دو نود x و y از درجه ۱ و درجه ۲ استفاده می‌شود.

همچنین به‌منظور مقایسه معیارهای شباهت فوق، از معیار دیگری که ترکیب دو معیار ۷ و ۸ است، استفاده شده تا بررسی شود در صورتی که هم اطلاعات همسایگان، هم‌درجه‌ی ۲ نود مذکور و هم اطلاعات مسیر بین دو نود در محاسبه‌ی احتمال ایجاد لینک دخالت داده شود، (DPNS) آنگاه با چه دقتی می‌توان شباهت بین دو نود را تخمین زد.

$$degree\_path\_neghibor\ Similarity\ (x,y) = x\_degree \times y\_degree \times \sum_{l=3}^{5} \alpha^{l}.|path_{x,y}^{l}| \times \sum_{Z\ \epsilon \Gamma(x)\cap \Gamma(y)} \frac{|C_{(Z)}|}{|\Gamma(z)|^{\beta C}} \qquad (۹)$$

که در این فرمول x_degree درجه‌ی نود x ، y_degree درجه‌ی نود y ، α ضریبی است که تأثیر مسیرهای به طول زیاد را در محاسبات تعیین می‌کند. $|path_{x,y}^{l}|$ بیانگر تعداد مسیرهای به طول l بین دو نود x و y است و $|C_{(Z)}|$ تعداد همسایگان مشترک z و x و y، $|\Gamma(z)|$ درجه‌ی نود z، C میانگین clustering coefficient در گراف و β یک مقدار ثابت است که با آزمایش روی دیتاست‌های مختلف مقدار بهینه‌ی آن از روش درون‌یابی به دست آمده است. در این معیار، از ترکیب اطلاعات درجه‌ی دو نود مورد نظر، تعداد همسایگان مشترک دو نود x و y از درجه ۱ و درجه ۲ و تعداد مسیرهای به طول مشخص l برای محاسبه‌ی احتمال ایجاد لینک بین دو نود x و y استفاده می‌شود.



شکل ۲: مثالی از گراف اجتماعی

به‌عنوان مثال در گراف شکل ۲ برای محاسبه‌ی شباهت بین دو نود x وy به دو روش فوق، مقادیر پارامترها در جدول ۲ نوشته شده است.

جدول ۲: مقادیر پارامترهای فرمول‌های ۷، ۸ و ۹ با توجه به شکل ۲

پارامترها	مقدار	مجموعه‌ی از نودها/مسیرها
$path_{x,y}^{1}$	∅	-------
$path_{x,y}^{2}$	2	{XCY,XDY}
$path_{x,y}^{3}$	3	{XACY,XBCY,XGDY}
$path_{x,y}^{4}$	3	{XFECY, XDECY, XCEDY}
$path_{x,y}^{5}$	∅	-----
y_degree	2	{C,D}
x_degree	6	{A,B,C,D,F,G}
$\Gamma(x)$	6	{A,B,C,D,F,G}
$\Gamma(y)$	2	{C,D}
Z	2	{C,D}
$C_{(z)}\ if\ z:c$	∅	-----
$C_{(z)}\ if\ z:d$	∅	-----
$\Gamma(z)\ if\ z:c$	6	{A,B,F,X,Y,E}
$\Gamma(z)\ if\ z:d$	4	{E,G,X,Y}

همان‌گونه که گفته شد در روش ارائه شده، محاسبات به‌صورت آنی صورت می‌گیرد و برای محاسبه‌ی شباهت دو نود بدون نیاز به اینکه مدلی از قبل‌ترین شده باشد، در همان زمان با محاسبه‌ی تعداد مسیرها و همسایگان مشترک بین دو نود می‌توانیم شباهت دو نود را حساب کنیم. این بحث از این جهت حائز اهمیت است که در صورتی که در شبکه تغییری رخ بدهد و نودی به شبکه اضافه یا حذف شود مشکلی در محاسبات ایجاد نمی‌شود زیرا روش پیشنهادی وابسته به ساختار شبکه نیست.

۴- نتایج

مشخصات دیتاست‌هایی که روش ارائه شده بر روی آن‌ها آزمایش شده در جدول ۳ آمده است که به ترتیب از ستون اول شامل: نام دیتاست، |V| تعداد نودها، |E| تعداد یال‌ها، K میانگین درجه،C میانگین clustering coefficient، ASPL[39] میانگین طول کوتاه‌ترین مسیر، D[40] قطر و H ناهمگنی در گراف موردنظر می‌باشد. تمامی این دیتاست‌ها جزء دیتاست‌های دنیای واقعی[41] هستند. همچنین مقدار Average clustering coefficient برای هر یک از دیتاست‌ها نیز در ستون ۵ جدول ۳ ذکر شده است.

جدول ۳: مشخصات دیتاست‌ها

network	|E|	K	|V|	C	D	H	ASPL
BUP	441	8.4	105	0.49	7	1.42	3.08
CEG	2148	14.46	297	0.29	5	1.80	2.46
UAL	2126	12.81	332	0.63	6	3.46	2.74
INF	2765	13.49	410	0.46	9	1.39	3.63
SMG	4916	9.6	1024	0.31	6	3.95	2.98
EML	5451	9.62	1133	0.22	8	1.94	3.61

CEG یک شبکه‌ی زیستی[۴۲]، SMG شبکه‌ی تألیف مشترک[۴۳]، UAL شبکه‌ی ترافیک فرودگاه، EML شبکه‌ی افرادی است که ایمیل برای یکدیگر فرستاده‌اند، BUP شبکه‌ی وبلاگ‌های سیاسی و INF شبکه‌ی ارتباطات رو در رو در یک نمایشگاه است. دیتاست‌های فوق در آدرس https://noesis.ikor.org/datasets/link-prediction موجود است.

آزمایشات بر روی سیستمی با RAM 32 و پردازنده‌ی اینتل core i5 با فرکانس‌های ۳.۱ و ۳.۴ گیگاهرتز انجام شد. مقداری که برای پارامترها در نظر گرفته شده به‌صورت زیر است:

$\alpha : 0.1 , 0.05$

$l : 3 - 5$

$\beta \; in \; AUC \; metric : 1.76$

$\beta \; in \; precision \; metric : 1.84$

مقدار α باید بین صفر و یک باشد. در این مقاله تنها دو مقدار ۰.۱ و ۰.۰۵ برای α به صورت آزمون و خطا در نظر گرفته شده است که جواب بهتری نسبت به سایر مقادیر امتحان شده دارد. برای شمارش مسیر بین دو نود باید $l > 2$ در نظر گرفته شود زیرا $l = 2$ ،یال مستقیم بین دو نود است. در این مقاله، مقدار l بین ۳ و ۵ در نظر گرفته شده است. یافتن تعداد مسیرهای به طول بیشتر علاوه بر صرف زمان خیلی زیاد، حاوی اطلاعات مفیدی برای پیش‌بینی لینک نمی‌باشد زیرا مسیرهای به طول کمتر مفیدتر هستند. مقادیر مناسب برای β نیز از طریق درون‌یابی بر روی دیتاست‌های مورد آزمایش به دست آمده است [۱۵].

به‌منظور ارزیابی عملکرد الگوریتم ارائه شده و الگوریتم‌های دیگر از دو معیار Area Under Curve (AUC) و precision استفاده می‌کنیم. شیوه‌ی محاسبه‌ی هر یک از دو معیار به‌صورت زیر است [۱۵]:

ابتدا یال‌ها در گراف موردنظر به روش five-folds cross validation به ۵ قسمت تقسیم می‌شود که یک قسمت به‌عنوان test data و ۴ قسمت به‌عنوان train data در نظر گرفته می‌شود. بنابراین train data شامل ۰/۸ از یال‌های گراف است و test data نیز شامل مابقی یال‌ها[۴۴] (۰/۲ از یال‌های گراف) به‌علاوه‌ی یال‌هایی که اصلاً وجود نداشته‌اند[۴۵] می‌باشد. برای نودهای دوسر یال‌ها در test data با معیار ارائه شده، شباهت حساب می‌شود و سپس یال‌ها بر اساس مقدار شباهت به صورت نزولی مرتب می‌شوند و سپس به تعداد ۰/۲ از یال‌های گراف اصلی از یال‌های مرتب‌شده‌ی test data برداشته می‌شود و به‌عنوان یال‌های پیش‌بینی شده به train data اضافه می‌شود. پس از آن true positive و false negative محاسبه شده و بر اساس آن میزان AUC و precision تخمین زده خواهد شد.

به‌منظور محاسبه‌ی مقدار AUC ، یک لینک از مجموعه‌ی test data (شامل لینک‌های non-existent و non-observation) و یک لینک از مجموعه‌ی non-existent (شامل ۰.۲ یال‌هایی که از گراف اصلی به‌عنوان non-existent برداشته شد) برمی‌داریم و مقدار AUC این دو لینک را که از رابطه‌ی زیر به دست آمده مقایسه می‌کنیم.

$$AUC = \frac{n_1 + 0.5n_2}{n} \qquad (۱۰)$$

در این رابطه n_1 تعداد دفعاتی است که مقدار AUC لینکی که از مجموعه‌ی test data برداشته شده است، بیشتر است، n_2 تعداد دفعاتی است که مقدار AUC لینکی که از مجموعه‌ی non-existent برداشته شده است، بیشتر است و n تعداد کل مقایسات است. در این مقاله n برابر با حاصل‌ضرب تعداد یال‌های هر دو مجموعه در هم است. درواقع تمام یال‌های هر دو مجموعه با هم مقایسه می‌شوند. معیار precision نیز از رابطه‌ی زیر به دست می‌آید.

$$precision = \frac{TP}{TP + FP} \qquad (۱۱)$$

که TP تعداد لینک‌هایی است که به‌درستی پیش‌بینی شده و $TP + FP$ تعداد کل لینک‌های پیش‌بینی شده است.

جدول ۴: بررسی نتایج الگوریتم Degree Neighbor Similarity (DNS) و الگوریتم‌های ADP، NCCCN، Triadic measure و CNDP بر روی ۶ دیتاست.

algorithm	Evaluation metric	BUP	CEG	UAL	INF	SMG	EML
ADP	precision	0.254	0.153	0.515	**0.419**	**0.159**	0.197
	AUC	0.749	0.800	**0.931**	0..869	**0.833**	0.797
NCCCN	precision	0.242	0.160	0.502	0.293	0.148	**0.211**
	AUC	0.753	0.781	0.919	0.879	0.812	0.802
Triadic Measure	precision	0.229	0.147	0.483	0.364	0.136	0.173
	AUC	0.733	0.734	0.893	0.483	0.782	0.785
CNDP	precision	0.263	0.150	0.478	0.434	0.143	0.203
	AUC	0.853	0.817	0.928	**0.920**	0.812	**0.820**
DNS	precision	**0.265**	**0.167**	**0.595**	0.368	0.142	0.151
	AUC	**0.855**	**0.821**	0.930	0.911	0.820	0.810

جدول ۵: بررسی نتایج الگوریتم Path Neighbor Similarity (PNS) و الگوریتم‌های ADP، NCCCN ، Triadic measure و CNDP بر روی ۶ دیتاست (α= ۰/۱, α= ۰/۰۵)

algorithm	Evaluation metric	BUP	CEG	UAL	INF	SMG	EML
ADP	precision	0.254	0.153	0.515	0.419	0.159	0.197
	AUC	0.749	0.800	0.931	0..869	0.833	0.797
NCCCN	precision	0.242	0.160	0.502	0.293	0.148	0.211
	AUC	0.753	0.781	0.919	0.879	0.812	0.802
Triadic Measure	precision	0.229	0.147	0.483	0.364	0.136	0.173
	AUC	0.733	0.734	0.893	0.483	0.782	0.785
CNDP	precision	**0.263**	0.150	0.478	0.434	0.143	0.203
	AUC	0.853	0.817	0.928	0.920	0.812	0.820
PNS **α= 0/1**	precision	0.258	**0.183**	0.589	0.401	0.159	0.194
	AUC	0.867	0.825	0.924	0.920	0.820	0.822
PNS **α= 0/05**	precision	0.256	0.175	**0.596**	**0.437**	**0.161**	**0.214**
	AUC	**0.870**	**0.837**	**0.931**	**0.924**	**0.836**	**0.824**

نتایج جدول ۴ حاکی از آن است که الگوریتم DNS که تلفیقی از اطلاعات همسایگان بین دو نود و درجه‌ی دو نود است بر روی دیتاست‌های BUP و CEG بهتر از الگوریتم‌های دیگر عمل می‌کند اما بر روی سایر دیتاست‌ها نتوانسته از رقیبان خود ببرد. بر روی دیتاهای INF و EML الگوریتم CNDP که تنها از اطلاعات همسایگان استفاده می‌کند و بر روی دیتای SMG الگوریتم CNDP که تنها از اطلاعات همسایگان استفاده می‌کند نتایج بهتری داشته است. می‌توان دریافت که درجه‌ی نودها بسیار تأثیرگذار نیست و نمی‌توان همواره ادعا کرد که هرچه نود درجه‌ی بیشتری داشته باشد احتمال تشکیل لینک توسط آن بیشتر می‌شود.

نتایج ارزیابی الگوریتم PNS بر روی دیتاست‌های فوق‌الذکر در جدول ۵ آمده است.. همان‌گونه که مشهود است، الگوریتم PNS که علاوه بر اطلاعات همسایگان مشترک دو نود، از اطلاعات مسیر بین دو نود نیز بهره می‌گیرد نتایج بهتری را نسبت به DNS داشته است. در جدول ۵ الگوریتم PNS با آلفای ۰.۱ صرف‌نظر از PNS با آلفای ۰.۰۵بر روی تمامی دیتاها حداقل در یک معیار AUC و یا precision بر الگوریتم‌های رقیب برتری داشته است. مشاهده

جدول ۶: بررسی نتایج الگوریتم‌های Degree Path Neighbor Similarity (DPNS) و الگوریتم‌های ADP، NCCCN ، Triadic measure و CNDP بر روی ۶ دیتاست (α= ۰/۱, α= ۰/۰۵)

algorithm	Evaluation metric	BUP	CEG	UAL	INF	SMG	EML
ADP	precision AUC	0.254 0.749	0.153 0.800	0.515 **0.931**	0.419 0..869	**0.159** 0.833	0.197 0.797
NCCCN	precision AUC	0.242 0.753	0.160 0.781	0.502 0.919	0.293 0.879	0.148 0.812	**0.211** 0.802
Triadic Measure	precision AUC	0.229 0.733	0.147 0.734	0.483 0.893	0.364 0.483	0.136 0.782	0.173 0.785
CNDP	precision AUC	0.263 0.853	0.150 0.817	0.478 0.928	**0.434** **0.920**	0.143 0.812	0.203 0.820
DPNS **α= 0/1**	precision AUC	0.267 0.858	0.166 0.819	0.602 0.921	0.384 0.913	0.142 0.818	0.161 0.820
DPNS **α= 0/05**	precision AUC	**0.272** **0.860**	**0.172** **0.826**	**0.606** 0.924	0.397 0.916	0.155 **0.834**	0.177 **0.821**

می‌شود که بر روی دیتاست‌های BUP، CEG، INF و EML از نظر معیار AUC به نسبت الگوریتم‌های دیگر برتری دارد و بر روی دیتاهای CEG، UAL و SMG از نظر معیار precision عملکرد بهتری داشته است.

با کم کردن مقدار α از ۰/۱ به ۰/۰۵ که موجب کمتر شدن تأثیر مسیرهای طولانی‌تر شده مشاهده می‌شود که الگوریتم ارائه شده بر روی تمامی دیتاست‌ها به جز BUP وCEG از نظر هر دو معیار precision و AUC عملکرد بهتری داشته است. و روی دیتاست BUP وCEG نیز از نظر AUC بر الگوریتم‌های دیگر برتری داشته است. البته قابل ذکر است که در CEG نیز الگوریتم PNS با آلفای ۰.۱مقدار Precision بالاتری به خود اختصاص داده است. از نتایج می‌توان متوجه شد که تأثیر اطلاعات مسیر بر روی بهبود دقت پیش‌بینی بسیار تأثیرگذار است و انتخاب مقدار مناسبα و تنظیم تأثیر مسیرهای طولانی می‌تواند دقت پیش‌بینی را به مقدار قابل توجهی افزایش دهد.

به‌منظور بررسی دقیق‌تر تأثیر اطلاعات مسیر و اطلاعات درجه‌ی نودها، با معیار شباهتی که در فرمول ۴ معرفی شده که از ترکیب هر دوی اطلاعات استفاده می‌کند، الگوریتم شناسایی لینک بر روی دیتاست‌های فوق انجام شد و نتایج در جدول ۶ و ۷ گزارش شده است.

جدول ۷: مقایسه الگوریتم‌های(Degree Neighbor Similarity (DNS و Path Neighbor Similarity (PNS) و Degree Path Neighbor Similarity (DPNS) به ازای دو مقدار ۰.۵ و ۰.۱ برای α بر روی ۶ دیتاست

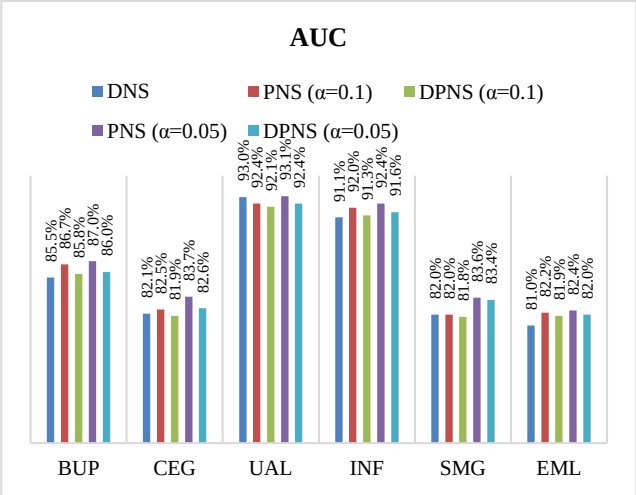
algorithm	α	Evaluation metric	BUP	CEG	UAL	INF	SMG	EML
DNS	—	precision AUC	0.265 0.855	0.167 0.821	0.595 0.930	0.368 0.911	0.142 0.820	0.151 0.810
PNS	0.1	precision AUC	0.258 0.867	**0.183** 0.825	0.589 0.924	0.401 0.920	0.159 0.820	0.194 0.822
DPNS	0.1	precision AUC	0.267 0.858	0.166 0.819	0.602 0.921	0.384 0.913	0.142 0.818	0.161 0.820
PNS	0.05	precision AUC	0.256 **0.870**	0.175 **0.837**	0.596 **0.931**	**0.437** **0.924**	**0.161** **0.836**	**0.214** **0.824**
DPNS	0.05	precision AUC	**0.272** 0.860	0.172 0.826	**0.606** 0.924	0.397 0.916	0.155 0.834	0.177 0.821

همان‌گونه که از نتیجه جدول ۶ مشاهده می‌گردد، ترکیب اطلاعات مسیر و درجه‌ی نودها از دقت PNS می‌کاهد و نتایج را بدتر از زمانی می‌کند که تنها از اطلاعات مسیر استفاده می‌شود. اما با این حال به نسبت الگوریتم‌های دیگر بهتر عمل می‌کند و مشاهده می‌شود که بر روی تمام دیتاست‌ها به جز INF در حداقل یکی از معیارها برتری دارد. در این جدول ارزیابی نتایج حاصل از DPNS با مقدار α=۰/۱ و مقدار α=۰/۰۵ برای معیار DPNS در نظر گرفته شده است. می‌توان مشاهده کرد که مقدار کمتر α همان‌گونه که در الگوریتم PNS موجب بهبود نتایج شده بود در DPNS نیز نتایج را بهبود می‌دهد و این بیانگر این است که تأثیر مسیرهای طولانی هرچه کمتر باشد دقت پیش‌بینی افزایش می‌یابد.

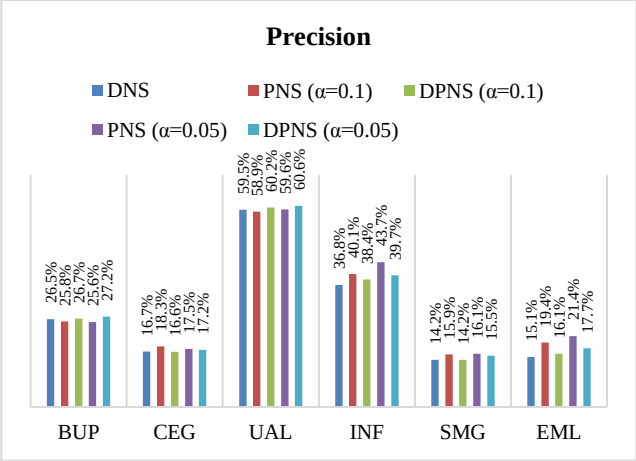
به‌منظور مقایسه‌ی بهتر معیارهای ارائه شده در این مقاله، در جدول ۷ نتایج هر یک از معیارها بر روی دیتاست‌های فوق‌الذکر گزارش شده است. برتری معیار PNS با مقدار α=۰/۰۵ به‌راحتی قابل مشاهده است. همان‌گونه که از نتایج در جدول ۷ برمی‌آید در مورد ۳ دیتاست INF، SMG و EML معیار PNS با α=۰/۰۵ عملکرد بهتر و دقت پیش‌بینی بالاتری داشته است. همچنین برای سایر دیتاها نیز می‌توان مشاهده نمود که از نظر مقدار AUC معیار PNS با α=۰/۰۵ بر سایر معیارها برتری دارد.

شکل‌های ۳ و ۴ جهت مقایسه‌ی بصری الگوریتم‌های ارائه شده در این مقاله، رسم شده‌اند. نمودار شکل ۳ عملکرد الگوریتم‌های فوق‌الذکر را بر اساس معیار AUC و نمودار شکل ۴ بر اساس معیار precision مقایسه می‌کند. همان‌گونه که مشاهده می‌شود در نمودار ۳ به ازای تمامی دیتاها الگوریتم PNS که از اطلاعات مسیر و همسایگان بهره می‌برد، بر اساس معیار AUC یا بهتر از سایر الگوریتم‌ها بوده یا نتیجه‌ای برابر با سایر الگوریتم‌ها داشته است، همچنین در نمودار شکل ۳ می‌توان مشاهده نمود که برای اکثر دیتاست‌ها الگوریتم DPNS با α=۰/۱، بدترین عملکرد را داشته است.در نمودار شکل ۴، برای دیتاست‌های INF و SMG و EML عملکرد بهتر PNS با مقدار α=۰/۰۵ به‌وضوح مشهود است. برای دو دیتاست BUP و UAL نیز DPNS، با α=۰/۰۵ ، با اختلاف اندکی بهتر از PNS عمل کرده است.

از مقایسه‌ی نتایج نمودارها و جداول می‌توان این‌گونه برداشت کرد که از بین معیارهای شباهت ارائه شده معیار PNS نتایج قابل قبولی بر روی شبکه‌های مورد بررسی داشته است. دقت پیش‌بینی این معیار برای شبکه‌های UAL و INF، SMG و EML بر اساس هر دو معیار Precision و AUC بهتر از سایر از الگوریتم‌های پیش‌بینی لینک بوده است و این نشان می‌دهد معیار ارائه شده برای شبکه‌های مورد بررسی کارآمد است.



شکل ۳: مقایسه الگوریتم‌های پیشنهادی با معیار AUC



شکل ۴: مقایسه الگوریتم‌های پیشنهادی با معیار Precision

۵- نتیجه‌گیری و کارهای آتی

در این مقاله سعی شد تا با ترکیب اطلاعات ساختاری گراف مانند اطلاعات مبتنی بر همسایگان نودها و اطلاعات مبتنی بر مسیر بین نودها و همچنین با بهره‌گیری از درجه‌ی نودها در اسنپ شات فعلی از شبکه، معیاری از شباهت معرفی شود که بتواند با دقت بهتری ایجاد یا عدم ایجاد لینک را در شبکه پیش‌بینی کند. به همین منظور سه معیار مختلف با ترکیب اطلاعات متفاوت از شبکه ارائه شد تا بتوان به‌صورت دقیق‌تری تأثیر هر یک از عوامل را در دقت پیش‌بینی لینک‌ها بررسی کرد. معیار اول degree-neighbor similarity (DNS) از ترکیب اطلاعات مربوط به درجه‌ی نودهای مبدأ و مقصد

و همچنین همسایگان مشترک مستقیم و درجه‌ی دوم دو نود به دست می‌آید. معیار دوم(Path-Neighbor Similarity (PNS از ترکیب اطلاعات مسیرهای به طول ۳ تا ۵ بین دو نود و همسایگان مشترک مستقیم و درجه دوم دو نود به دست می‌آید. معیار سوم نیز Degree-Path-Neighbor Similarity (DPNS) از ترکیب اطلاعات درجه‌ی دو نود مذکور، مسیرهای بین دو نود و همسایگان مستقیم و درجه دوم دو نود به دست می‌آید.

نتایج حاکی از آن است DNS بر روی بعضی از دیتاها مانند BUP و CEG بهتر از سایر الگوریتم‌ها عمل کرده اما بر روی بقیه‌ی دیتاها موفق نبوده است. نتایج الگوریتم‌های PNS و DPNS با مقادیر مختلف α نشان می‌دهد که این الگوریتم بهتر از DNS عمل می‌کند و به ازای α=۰/۰۵ بر روی تمام دیتاست‌ها نتایج بهتری را نسبت به رقیبان دارد. همچنین می‌توان مشاهده کرد که PNS نسبت به DPNS به ازای مقادیر مختلف α ، عملکرد بهتری از خود نشان داده و این گویای این است بهره‌گیری از اطلاعات مسیر بین دو نود و همسایگان مشترک دو نود بسیار حائز اهمیت است در حالی که درجه‌ی نودها تأثیر چندانی بر روی دقت پیش‌بینی لینک‌ها ندارد.

از جمله محدودیت‌های روش ارائه شده پیچیدگی زمانی زیاد برای اجرای الگوریتم می‌باشد. یافتن همسایگان مشترک درجه‌ی دوم دو نود مورد نظر دارای پیچیدگی زمانی $O(N^3)$ است. همچنین پیدا کردن مسیرهای به طول l بین دو نود نیز زمان بسیار زیادی تحمیل می‌کند، به همین دلیل، برای اجرای الگوریتم در شبکه‌های بزرگ زمان خیلی زیادی صرف می‌شود و از آنجا که این روش ترکیبی از چند روش است زمان بیشتری نسبت به روش‌های مشابه صرف اجرای الگوریتم می‌شود.

برای کارهای آتی، همان‌گونه که در بررسی کارهای پیشین نیز اشاره شد، بهره‌گیری از اطلاعات خوشه‌ها در شبکه می‌تواند در نتیجه‌ی کار بهبود ایجاد کند. همچنین در شبکه‌هایی که اطلاعات اختصاصی مربوط به نودها نیز موجود است می‌توان از این اطلاعات، در جهت بهبود دقت پیش‌بینی بهره گرفت.

مراجع

[1] Borgatti, S. P., Everett, M. G., & Johnson, J. C. (2018). *Analyzing social networks.* Sage.

[2] *Wasserman, Stanley; Faust, Katherine (1994). "Social Network Analysis in the Social and Behavioral Sciences". Social Network Analysis: Methods and Applications. Cambridge University Press. pp. 127. ISBN 9780521387071.*

[3] Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American society for information science and technology*, *58*(7), 1019-1031.

[4] Aiello L M, Barrat A, Schifanella R, et al. Friendship prediction and homophily in social media. ACM Trans Web, 2012, 6: 9

[5] Mori J, Kajikawa Y, Kashima H, et al. Machine learning approach for finding business partners and building reciprocal relationships. Expert Syst Appl, 2012, 39: 10402–10407

[6] Wohlfarth T, Ichise R. Semantic and event-based approach for link prediction. In: Proceedings of the 7th International Conference on Practical Aspects of Knowledge Management (PAKM'08), Yokohama, 2008. 50–61

[7] Raeder T, Lizardo O, Hachen D, et al. Predictors of short-term decay of cell phone contacts in a large scale communication network. Soc Netw, 2011, 33: 245–257

[8] Marchette D J, Priebe C E. Predicting unobserved links in incompletely observed networks. Comput Stat Data Anal, 2008, 52: 1373–1386

[9] Kim M, Leskovec J. The network completion problem: inferring missing nodes and edges in networks. In: Proceedings of the 11th SIAM International Conference on Data Mining (SDM'11), Mesa, 2011. 47–58

[10] Barab´asi A L, Jeong H, N´eda Z, et al. Evolution of the social network of scientific collaborations. Physica A, 2002, 311: 590–614

[11] Almansoori W, Gao S, Jarada T N, et al. Link prediction and classification in social networks and its application in healthcare and systems biology. Netw Model Anal Health Inform Bioinform, 2012, 1: 27–36

[12] Wang, P., Xu, B., Wu, Y., & Zhou, X. (2015). Link prediction in social networks: the state-of-the-art. *Science China Information Sciences*, *58*(1), 1-38.

[13] Adamic L A, Adar E. Friend and neighbors on the web. Soc Networks, 2003, 25: 211–230

[14] Katz L. A new status index derived from sociometric analysis. Psychometrika, 1953, 18: 39–43

[15] Rafiee, Samira, Chiman Salavati, and Alireza Abdollahpouri. "CNDP: Link prediction based on common neighbors degree penalization." *Physica A: Statistical Mechanics and its Applications* 539 (2020): 122950.

[16] V. Martínez, F. Berzal, J.-C. Cubero, Adaptive degree penalization for link prediction, J. Comput. Sci. 13 (2016) 1–9.

[17] F. Li, J. He, G. Huang, Y. Zhang, Y. Shi, R. Zhou, Node-coupling clustering approaches for link prediction, Knowl.-Based Syst. 89 (2015) 669–680.

[18] F. Aghabozorgi, M.R. Khayyambashi, A new similarity measure for link prediction based on local structures in social networks, Physica A 501 (2018) 12–23.

[19] Yu, Chuanming, et al. "Similarity-based link prediction in social networks: A path and node combined approach." *Journal of Information Science* 43.5 (2017): 683-695.

[20] Yin Z, Gupta M, Weninger T and Han J. LINKREC: A unified framework for link recommendation with user attributes and graph structure. In: Proceedings of the 19th International Conference on World Wide Web. North Carolina, USA: ACM, 2010, pp. 1211–1212.

[21] Wang D, Pedreschi D, Song C, Giannotti F and Barabasi AL. Human mobility, social ties, and link prediction. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '11. New York: ACM, 2011, pp. 1100–1108.

[22] Z. Liu, Q.M. Zhang, L. L, T. Zhou, Link prediction in complex networks: a local naive bayes model, EPL Europhys. Lett. (2011) 96.

[23] J. Feng, J. Zhao, K. Xu, Link prediction in complex networks: a clustering perspective, Eur. Phys. J. B 85 (2012) 1–9.

[24] Bastami, E., Mahabadi, A., & Taghizadeh, E. (2019). A gravitation-based link prediction approach in social networks. *Swarm and evolutionary computation*, *44*, 176-186.

[25] Newman M E J. Clustering and preferential attachment in growing networks. Phys Rev E, 2001, 64: 025102

[26] Zhou, M., Li, B., Yang, M. & Pan, L. TeleGraph: A Benchmark Dataset for Hierarchical Link Prediction. *arXiv preprint arXiv:2204.07703* (2022).

[27] Long, Y. *et al*. Pre-training graph neural networks for link prediction in biomedical networks. *Bioinformatics* **38**, 2254–2262 (2022).

[28] Kumar, S., Mallik, A. & Panda, B. S. Link prediction in complex networks using node centrality and light gradient boosting machine. *World Wide Web* 1–27 (2022).

پاورقی‌ها:

[1] Link prediction
[2] Neighbor-based
[3] Path-based
[4] Graph topology
[5] networks
[6] Social systems
[7] attributes
[8] Social structure
[9] link
[10] actors
[11] Link prediction
[12] Recommender systems
[13] e-commerce
[14] Bioinformatics
[15] biology
[16] Gene expression
[17] node-based
[18] topology-based
[19] social theory-based
[20] topology-based
[21] Neighbor-based
[22] path-based
[23] random walk-based
[24] regression
[25] Adaptive Degree Penalization
[26] penalized
[27] motif
[28] Bastami et.al.
[29] Gravity-based
[30] Yin et. al.
[31] Wang et. al.
[32] mobility
[33] Preprocess
[34] directed
[35] undirected
[36] Simple graph
[37] Computational complexity
[38] regression
[39] Average shortest path
[40] diameter
[41] Real-world datasets
[42] biological
[43] Co-authorship
[44] Non-existent
[45] Non-observation