سال سیزدهم، شماره‌های ۴۷ و ۴۸، بهار و تابستان ۱۴۰۰
صص: ۱۷۸_۱۸۵

دو فصلنامه علمی
فناوری اطلاعات و ارتباطات ایران

یک روش پیش‌بینی پیوند مبتنی بر همسایه برای شبکه دوبخشی

گلشن سندسی* سید علیرضا هاشمی گلپایگانی علیرضا صائبی*****

*دانش‌آموخته کارشناسی ارشد، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر

** استادیار، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر

***دانش‌آموخته کارشناسی ارشد، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر

تاریخ دریافت: ۱۳۹۹/۰۸/۲۱ تاریخ پذیرش: ۱۳۹۹/۱۲/۱۹

نوع مقاله: پژوهشی

چکیده

پیش‌بینی پیوند، یکی از روش‌های تحلیل شبکه اجتماعی است. شبکه های دوبخشی یکی از انواع شبکه های پیچیده هستند که بسیاری از وقایع طبیعی، با استفاده از آن قابل مدل شدن هستند. در این مقاله، روشی برای پیش‌بینی پیوند در شبکه دوبخشی ارائه شده‌است. با توجه به اینکه روش‌های پیش‌بینی پیوند در شبکه یک بخشی برای استفاده در شبکه دوبخشی کارایی پایینی دارند و کارآمد نیستند، نیاز است برای حل این مسئله از روش‌هایی مختص شبکه دوبخشی استفاده شود. هدف این پژوهش، ارائه روشی جدید، متمرکز و جامع مبتنی بر همسایه است، که عملکردی بهتر از روش‌های کلاسیک موجود داشته باشد. روش پیشنهادی از ترکیب معیارهایی بر اساس همسایگی تشکیل شده‌است. معیارهای کلاسیک پیش‌بینی پیوند با اعمال تغییراتی برای شبکه دوبخشی تعریف شده‌اند. این معیارهای تغییر یافته، ارکان اصلی معیار پیشنهادی را تشکیل می‌دهند. این روش علاوه بر سادگی و پیچیدگی پایین، از کارایی بالایی برخوردار است و روش‌های کلاسیک مبتنی بر همسایه را در مجموعه داده‌های مورد بررسی به طور میانگین بیش از ۱۵٪ بهبود داده است.

واژگان کلیدی: نظریه گراف، تحلیل شبکه اجتماعی، شبکه دوبخشی، پیش‌بینی پیوند، پیش‌بینی پیوند در شبکه دوبخشی

۱ مقدمه

استفاده از شبکه‌ها و گراف‌ها، یکی از روش‌های مدل سازی مسائل دنیای واقعی به منظور حل آنها است. شبکه‌ها به خودی خود، می‌توانند وقایع بسیاری را در ساختارهای اجتماعی تو صیف کنند. تحلیل شبکه اجتماعی یکی از روش‌های تحلیل شبکه است که در سال‌های اخیر، در حوزه‌های تحقیقاتی مختلف بسیار مورد توجه قرار گرفته است. تحلیل شبکه اجتماعی یک انتزاع از روابط اجتماعی در دنیای واقعی است و به وسیله آن می‌توان افراد و روابط بین آنها را در یک زمان خاص به صورت یک گراف ایستا و یا در چند زمان به صورت یک گراف پویا نمایش داد. تحلیل شبکه اجتماعی از نظریه گراف برای ساخت این انتزاع استفاده می‌کند؛ به طوری که در آن گره‌ها افراد و یال‌ها روابط بین آنهاست[۱]. این

نویسنده مسئول: سید علیرضا هاشمی گلپایگانی

sa.hashemi@aut.ac.ir

ارتباطات می‌توانند صریح و یا ضمنی باشند. با توجه به اینکه پایه تشکیل این شبکه نظریه گراف است، می‌توان از معیارهای بر پایه گراف مانند مرکزیت، چگالی، قطر، ماژولاریتی و غیره بهره گرفت[۲]. مصورسازی[۱]، تحلیل ساختاری شبکه[۲]، تشخیص جوامع[۳]، شبکه‌های جهان کوچک[۴]، تحلیل موتیف‌ها[۵]، پیش‌بینی پیوند[۶] و انتشار[۷] از جمله روش‌های مبتنی بر تحلیل شبکه اجتماعی هستند. از زمینه های به کارگیری تحلیل شبکه اجتماعی می‌توان به کسب‌وکارهای اینترنتی و بازاریابی الکترونیکی، علوم اجتماعی، توصیه‌گری، کشف تقلب و حوزه علوم سلامت اشاره کرد.

شبکه دوبخشی، یکی از انواع مهم شبکه‌های پیچیده است که در آن گره‌ها به دو بخش تقسیم می‌شوند. در این شبکه‌ها یال‌ها گره‌های بخش‌ها مختلف را به هم متصل می‌کند و هیچ یالی بین دو گره از یک بخش وجود ندارد[۳]. بسیاری از شبکه‌های دنیای واقعی در اصل شبکه‌های دوبخشی هستند، مانند افراد و اقلام خریداری شده، افراد و بیماری‌ها، بیماری‌ها و ژن‌ها، مقالات و نویسندگان، کلمات و متون، سرمایه‌گذاران و شرکت‌ها.

پیش‌بینی پیوند یکی از مسایل پراهمیت در ارتباط با شبکه و یکی از روش‌های اصلی در تحلیل شبکه اجتماعی است. در شبکه اجتماعی $G(V,E)$ در زمان t، که V و E مجموعه گره‌ها و یال‌ها هستند، پیش‌بینی پیوند به شناسایی یال‌های جدید، یال‌های حذف‌شده یا یال‌های مشاهده نشده در زمان t' می‌پردازد، به طوری که $t'>t$ باشد. برای حل مسئله پیش‌بینی پیوند، لازم است آرایش یا تجزیه احتمالات وجود یال بین همه گره‌ها بررسی شود که معمولا از شباهت سنجی برای این کار استفاده می شود. روش‌های پیش‌بینی پیوند به این ترتیب دسته‌بندی می‌شوند[۴]:

الف) روش‌های مبتنی بر گره

ب) روش های مبتنی بر توپولوژی: مبتنی بر همسایه، مبتنی بر مسیر[۸] و مبتنی بر قدم‌زنی تصادفی[۹].

ج) روش‌های مبتنی بر نظریه اجتماعی: مبتنی بر جوامع، مبتنی بر ۳تایی‌ها، مبتنی بر حفره‌های ساختاری، مبتنی بر استحکام پیوند و مبتنی بر هموفیلی.

شاخص‌های پیش‌بینی پیوند عمدتا مبتنی بر توپولوژی هستند و به ۳ گروه اصلی تقسیم می‌شوند[۱] و[۲]:

۱) معیارهای مبتنی بر همسایه‌ها، مانند همسایه‌های مشترک[۱۰]، ضریب ژاکارد[۱۱]، ضریب آدامیک-آدار[۱۲]، انضمام ترجیحی[۱۳] و تخصیص منابع[۱۴].

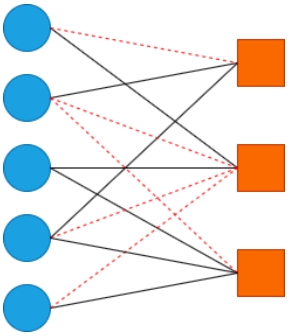
۲) معیارهای مبتنی بر مسیر، مانند مسیر محلی[۱۵]، معیار کتز[۱۶] و شباهت استحکام ارتباط[۱۷].

۳) معیارهای مبتنی بر قدم‌زنی تصادفی، مانند زمان برخورد[۱۸]، زمان طی کردن مسیر[۱۹]، زمان شباهت کسینوسی[۲۰] و رتبه صفحه ریشه‌دار[۲۱].

شکل ۱ شمایی کلی از یک شبکه دوبخشی ارائه می‌دهد. در این شکل، دو نوع گره وجود دارد(دایره و مربع). یال های مشکی، یال‌های موجود در شبکه هستند. یال‌های قرمز و خط‌چین، یال‌هایی هستند که پیش‌بینی پیوند، قصد دارد به وجود آمدن آنها را پیش‌بینی کند.

---

[۱] Visualization
[۲] Structural analysis
[۳] Community detection
[۴] Small-world networks
[۵] Motif analysis
[۶] Link Prediction
[۷] Diffusion
[۸] Path
[۹] Random walk
[۱۰] Common Neighbors(CN)
[۱۱] Jaccard's Coefficient(JC)
[۱۲] Adamic-Adar Coefficient(AA)
[۱۳] Preferential Attachment(PA)
[۱۴] Resource Allocation(RA)
[۱۵] Local Path(LP)
[۱۶] Katz
[۱۷] Relation Strength Similarity(RSS)
[۱۸] Hitting Time(HT)
[۱۹] Commute Time(CT)
[۲۰] Cosine Similarity Time(CST)
[۲۱] Rooted PageRank(RPR)



شکل ۱. شمایی از پیش‌بینی پیوند در شبکه دوبخشی

بسیاری از مسائل دنیای واقعی با استفاده از شبکه‌ها قابل مدل‌سازی هستند، تحلیل شبکه اجتماعی رویکردی مناسب برای تحلیل این شبکه‌هاست. پیش‌بینی پیوند یکی از روش‌های پراهمیت در تحلیل شبکه اجتماعی است و در زمینه‌های مختلفی مانند انواع توصیه‌گری و پیش‌بینی تقاضا، پیش‌بینی روابط دوطرفه، تکمیل و رشد شبکه، پیش‌بینی پیوندهای اجتماعی مانند دوستی و روابط کاری و هم-نویسندگی[۱] در کارهای علمی، پیش‌بینی عوارض جانبی داروها، مطالعه ژن‌ها، شناسایی جوامع ناهنجار کاربرد دارد[۴]و[۵]. مسئله پیش‌بینی پیوند به طور عمده درمورد شبکه‌های یک بخشی مطرح می‌شود. بنابراین روش مناسب جهت پیش‌بینی پیوند در شبکه دوبخشی، نیاز به مطالعه و بررسی دقیق‌تر دارد. با توجه به زمینه کاربرد، بهتر است روش پیش‌بینی پیوندی انتخاب شود که معیارهای استفاده‌شده در آن، با زمینه مسئله منطبق باشد.

هدف این مقاله، ارائه یک روش پیش‌بینی پیوند مخصوص شبکه‌های دوبخشی است که شبکه را از دیدگاه‌های متفاوت مبتنی بر همسایه مورد بررسی قرار دهد و دارای عملکردی مطلوب باشد. در ادامه، در بخش ۲ به مرور مطالعات انجام‌شده در زمینه‌های مرتبط پرداخته شده‌است. در بخش ۳، روش پیشنهادی ارائه و توضیح داده شده‌است. در بخش ۴، داده‌های مورد استفاده و نحوه پیاده‌سازی بررسی شده‌است و بخش ۵ به بیان روش ارزیابی و نتایج آن پرداخته است.

۲ مطالعات پیشین

در این بخش به مرور مطالعات انجام شده درباره پیش‌بینی پیوند در شبکه دوبخشی پرداخته شده است. یکی از روش های پایه برای پیش‌بینی پیوند در گراف دوبخشی، پیش‌بینی پیوند در گراف یک بخشی نگاشت‌شده گراف دوبخشی[۲] است.[۳] در گراف نگاشت‌شده، با استفاده از نوعی وزن دهی، پیش‌بینی پیوند را انجام می دهد. [۶] پیوندهای داخلی را به گراف نگاشت‌شده اضافه و سپس با استفاده از معیار های پیش‌بینی پیوند، پیوند های آتی را پیش‌بینی کرده است. در پژوهش [۷]، مبتنی بر روش[۶]، پس از دسته‌بندی یک نوع از گره ها که درمورد آن اطلاعات موجود است و اضافه یال بین گره‌های داخل هر دسته، اقدام به پیش‌بینی پیوند کرده است. [۸] پس از نگاشت گراف دوبخشی به گراف وزن‌دار یک بخشی، درجه اهمیتی برای یال‌های پرامتیاز درنظر گرفته‌است و بررسی تغییرات درگذر زمان، به پیش‌بینی پیوندهای جدید در شبکه‌های پویای پیچیده[۳] پرداخته است. نویسندگان [۸] در [۹] نیز با درنظرگرفتن اینکه تغییرات شبکه در چه زمانی رخ داده‌اند، آستانه تعریف پیوندهای قوی و ضعیف را مورد مطالعه قرار داده‌اند.

گروهی از پژوهش‌های انجام شده، روش‌هایی مبتنی بر شباهت ارائه داده‌اند و معیارهای کلاسیک پیش‌بینی پیوند را با جایگزینی همسایه با همسایه همسایه گره‌ها، با شبکه دوبخشی منطبق کرده‌اند[۴]،[۱۰] و[۱۱]. [۱۲] با استفاده از معیارهای کلاسیک پیش‌بینی پیوند، به پیش‌بینی ارتباطات ناموجود در شبکه‌های آشیانه‌ای[۴] پرداخته است. [۱۳] از ترکیب انضمام ترجیحی و خوشه‌بندی استفاده کرده است و پیوندها را داخل هر بخش پیش‌بینی می‌کند. [۱۴] یک روش از ترکیب ساختار مبتنی بر شباهت و مدل ویژگی‌های پنهان ارائه کرده‌است. [۱۵] با وزن‌دهی بر پایه دانش زمینه‌ای مسئله و ترکیب آن با تخصیص منابع، و [۱۶] با تبدیل ماتریس مجاورت بر اساس شباهت بین گره‌های دو گروه، پیش‌بینی پیوند را انجام داده‌است. درروش مورد استفاده در[۱۶]و[۱۷]، CORLP[۵]، ایجاد یال با مفاهیم مثبت و منفی درنظر گرفته شده است. به این معنی که پسندیدن یا نپسندیدن یک قلم، به طور جداگانه لحاظ شده است[۱۸]. [۱۹] با استفاده از اندیس های اطلاعات محلی[۶]، روش های اطلاعات مسیر و دانش زمینه‌ای روشی ارائه کرده‌است. [۲۰] نیز بر مبنای دانش زمینه‌ای درمورد بخشی از گره‌ها روش خود را ارائه کرده‌است. [۲۱] با ترکیب معیارهای پیش‌بینی پیوند و استفاده از شناسایی جوامع، روش خود را برای پیش‌بینی پیوند در گراف‌های چند لایه معرفی کرده‌است.

در پیش‌بینی پیوند شبکه یک بخشی، معمولا دو حالت کامل کردن مثلث‌ها و خوشه‌بندی درنظر گرفته می‌شوند. این فرضیات در شبکه دوبخشی درست نیستند و درصورت استفاده از روش‌های پیش‌بینی پیوند شبکه یک بخشی، در شبکه دوبخشی، کارایی قابل قبولی

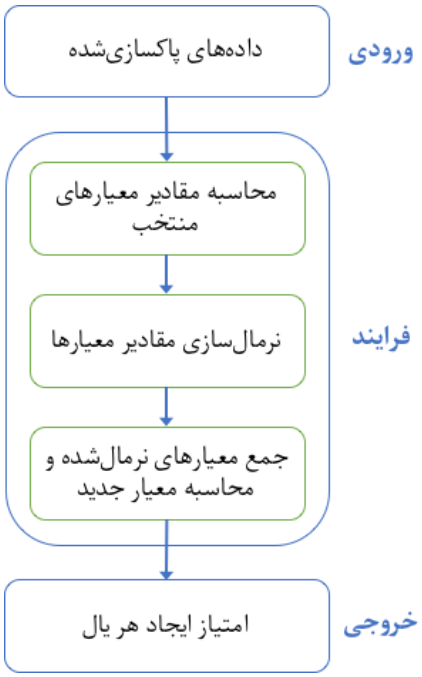
---

[۱] Co-authorship
[۲] Projected graph
[۳] Dynamic complex networks
[۴] Nested networks
[۵] Complex Representation-based Link Prediction
[۶] Local information index

حاصل نخواهد شد[۴]. از جهتی، از آنجایی که بسیاری از موضوعات دنیای واقعی در اصل یک شبکه دوبخشی هستند، بهتر است به منظور جلوگیری از دست رفتن اطلاعات، شبکه دوبخشی به شبکه یک بخشی تبدیل نشود. پیش‌بینی پیوند شبکه‌های دوبخشی، در پژوهش‌های بسیار کمتری مورد مطالعه قرار گرفته اند و مدل های ارائه شده، با فرضیات محدود و یا در کنار دیگر روش‌ها ارائه شده‌اند. موضوع دیگر، استفاده از معیارهای مختلف پیش‌بینی پیوند است که در بعضی مسائل، برخی از این معیارها دارای مفهوم نیستند و بنابراین استفاده از آنها درست نیست. گروهی دیگر از این روش‌ها پیچیدگی بالایی دارند و به سادگی قابل به‌کارگیری نیستند. همچنین در پژوهش‌های گذشته، روشی که روی همسایگی و معیارهای مبتنی بر همسایه متمرکز باشد و مسئله را از جنبه‌های مختلف بررسی کند، مشاهده نشده‌است. در این راستا، این پژوهش به ارائه یک روش پیش‌بینی پیوند جامع، با پیچیدگی پایین و مبتنی بر همسایه با تمرکز بر شبکه‌های دوبخشی ارائه داده است که گره‌های پراهمیت آن، گره‌هایی هستند که از اهمیت بالایی برخوردارند و داده‌های آن دارای توزیعی مشابه توزیع Power Law هستند.

۳ روش پیشنهادی

ایده روش پیشنهادی، استفاده از معیارهای شناخته‌شده و کلاسیک پیش‌بینی پیوند است که برای انطباق با شبکه دوبخشی تغییر یافته و متناسب با شبکه دوبخشی تعریف شده‌اند. با توجه به اهمیت معیارهای پایه، معیار تعریف‌شده جدید به ترکیب معیارهای کلاسیک می‌پردازد. در این روش به دلیل جامعیت و دامنه استفاده بیشتر، تمرکز روی معیارهای مبتنی بر همسایه است. معیارهای مبتنی بر همسایه، اصلی‌ترین گروه معیارهای پیش‌بینی پیوند هستند و در اکثر مسائل پیش‌بینی پیوند، به نوعی مورد استفاده قرار می‌گیرند. به منظور بهره‌گیری از دیدگاه‌های متفاوت به دست‌آمده با استفاده از این معیارها، تعریف و محاسبه یک معیار جامع و جدید مبتنی بر همسایه پیشنهاد می‌شود. تمرکز اصلی روش پیشنهادی، بر اساس اهمیت گره‌های پردرجه است. در این راستا، از میان معیارهای گروه مبتنی بر همسایه، شاخص‌هایی مورد استفاده قرار گرفته‌اند که امتیاز بالاتری را به گره‌های با درجه بالاتر اختصاص می‌دهند. در کنار این شاخص ها، شاخص های شباهت‌سنجی نیز به منظور محاسبه شباهت بین گره‌ها از دیدگاه‌های مختلف و بدون درنظر گرفتن درجه گره‌ها استفاده شده‌است. شاخص‌های مورد نظر در این پژوهش، ضریب ژاکارد، شباهت کسینوسی سالتون[۱]، انضمام ترجیحی، Hub Depressed و لیخت-هولم-نرمن[۲] هستند. معیار های نام‌برده، همگی از نوع معیارهای بر پایه همسایه هستند. بین این معیارها، ضرایب ژاکارد و شباهت کسینوسی سالتون، شباهت بین دو گره را محاسبه می‌کنند. با توجه به اهمیت درجه گره‌ها در شبکه‌هایی که مدل رشد آنها از قانون انضمام ترجیحی پیروی می‌کند، معیارهای انضمام ترجیحی، HD و LHN برای به کارگیری در معیار جدید انتخاب شده‌اند. در این معیارها، گره های پردرجه از اهمیت بیشتری برخوردارند. بسیاری از اتفاقات طبیعی از قاعده انضمام ترجیحی پیروی می‌کنند. با وجود اینکه تمامی معیارهای نام‌برده برای هر نوع توزیع داده قابل به کارگیری هستند، با توجه به اهمیت گره‌های پردرجه در اکثر این معیارها، به نظر می‌رسد به منظور استفاده در توزیع های لگاریتمی و کشیده به راست[۳] مناسب‌تر باشند. فرایند کلی این پژوهش در نمودارمفهومی شکل ۲ نشان داده شده است.



شکل ۲. نمودار مفهومی روش پیشنهادی

با توجه به اینکه بخشی از اطلاعات ضمنی که در برخی موارد بسیار مهم هستند، در تبدیل شبکه دوبخشی به یک بخشی از دست می‌رود[۲۲]، از استفاده از گراف نگاشت‌شده امتناع شده و به تعریف مجدد معیارهای پیش‌بینی پیوند روی شبکه دوبخشی پرداخته شده‌است. این معیارها در شبکه دوبخشی با جایگزینی «همسایه» با «همسایه همسایه» در ادامه تعریف می‌شوند. در این تعاریف، $\Gamma(x)$ مجموعه همسایه های گره x، $\Gamma'(x)$ مجموعه همسایه های

[۱] Salton Cosine Similarity(SC)
[۲] Leicht-Holme-Nerman(LHN)
[۳] Right-skewed

همسایه‌های گره x و $|\Gamma(x)|$ تعداد همسایه‌های گره x را نشان می‌دهد.

ضریب ژاکارد(JC): ضریب ژاکارد، یک معیار مبتنی بر شباهت است که تعداد همسایه‌های مشترک را نرمال‌سازی می‌کند و بنابراین، اطلاعات مفیدتری به نسبت تعداد همسایه های مشترک ارائه می‌دهد[۲۴]. این شاخص، ارزش بالاتری برای زوج گره‌هایی درنظر می‌گیرد، که نسبت تعداد همسایه‌های مشترک آنها به کل تعداد همسایه‌های آنها، مقدار بیشتری باشد. این معیار جزو اصلی‌ترین معیارهای مورد استفاده به منظور شباهت‌سنجی است و به طور خاص در پیش‌بینی پیوند به کار گرفته می‌شود. ضریب ژاکارد در شبکه دوبخشی با استفاده از رابطه ۱ محاسبه می‌شود:

$$JC(x,y)=\frac{|\Gamma(x)\cap\Gamma'(y)|}{|\Gamma(x)\cup\Gamma'(y)|} \quad (۱)$$

شباهت کسینوسی سالتون(SC): این شاخص یک معیار کسینوسی جامع، پرکاربرد و پایه در مبحث پیش‌بینی پیوند است که مانند ضریب ژاکارد، به اندازه‌گیری میزان شباهت بین دو گره می‌پردازد و با استفاده از رابطه ۲ محاسبه می‌شود:

$$SC(x,y)=\frac{|\Gamma(x)\cap\Gamma'(y)|}{\sqrt{|\Gamma(x).\Gamma'(y)|}} \quad (۲)$$

انضمام ترجیحی(PA): این معیار بیان می‌کند که یال‌های جدید، با احتمال بالاتری به گره‌هایی متصل می‌شوند که درجه بالاتری دارند و این احتمال متناسب با اندازه همسایگی گره است[۲۴]. ضرب تعداد همسایه‌های دو گره، اندازه همسایگی آنها را مشخص می‌کند و با رابطه ۳ محاسبه می‌شود:

$$PA(x,y)=|\Gamma(x)| \, . \, |\Gamma(y)| \quad (۳)$$

Hub Depressed (HD): با درنظر داشتن گره‌های پر درجه، این معیار امتیاز بین یک زوج گره را هم‌پوشانی توپولوژیک دو گره تعریف می‌کند[۲۵] که از طریق رابطه ۴ محاسبه می‌گردد:

$$HD(x,y)=\frac{|\Gamma(x)\cap\Gamma'(y)|}{\max(|\Gamma(x)| \, , \, |\Gamma'(y)|)} \quad (۴)$$

لیخت-هولم-نرمن(LHN): این معیار، ارزش بیشتری به زوج گره‌هایی اختصاص می‌دهد که تعداد همسایه‌های مشترک بیشتری به نسبت تعداد مورد انتظار این همسایه‌ها دارند(رابطه ۵)[۲۶]:

$$LHN(x,y)=\frac{|\Gamma(x)\cap\Gamma'(y)|}{|\Gamma(x).\Gamma'(y)|} \quad (۵)$$

با توجه به دامنه متفاوت و اختلاف مقادیر این معیارها، به منظور اثرگذاری یکسان در محاسبه معیار جدید، مقادیر پارامترهای برگزیده بین ۰ و ۱ با استفاده از مقیاس کردن ویژگی[۱] یا نرمال‌سازی کمینه و بیشینه (رابطه ۶)، نرمال‌سازی می‌شوند.

$$X'=\frac{X-X_{min}}{X_{max}-X_{min}} \quad (۶)$$

معیار پیشنهادی برای پیش‌بینی پیوند در شبکه دوبخشی تحت عنوان ضریب مبتنی بر همسایه[۲](NBC)، به منظور استفاده از اطلاعات به دست‌آمده از هر معیار با توجه به مفاهیم متفاوت آنها، استفاده از جمع مقادیر نرمال شده امتیازات این معیارها را پیشنهاد می‌کند. این معیار با استفاده از رابطه ۷ محاسبه می شود که در آن JC’ ، PA’ ، SC’ ، HD’ و LHN’ مقادیر نرمال‌شده این معیارها هستند.

$$NBC(x,y)=JC'(x,y)+PA'(x,y)+SC'(x,y)+HD'(x,y)+LHN'(x,y) \quad (۷)$$

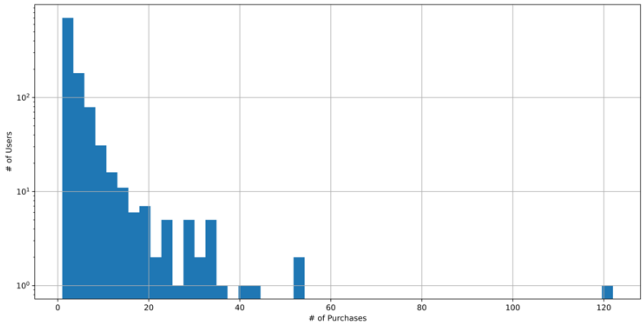
۴ تحلیل داده و شبکه

در این پژوهش از دو سری مجموعه داده فروشگاهی استفاده شده‌است. مجموعه اول مربوط به فروشگاه و شبکه اجتماعی آریو[۳] است که یک وب سایت داخلی خرید و فروش بازی‌های کامپیوتری است که از طرف این فروشگاه در دسترس نویسندگان قرار گرفته‌است. مجموعه دوم مربوط به یک فروشگاه خرده‌فروشی آنلاین است. این مجموعه داده از انبار یادگیری ماشین دانشگاه ارواین کالیفرنیا[۴] برداشته شده و برای عموم در اینجا[۵] قابل دسترسی است. در پیش‌پردازش هر دو مجموعه داده، داده‌های پرت و خارج از محدوده[۶] شناسایی و حذف شده‌اند. داده‌های پرت در هر دو مورد، داده‌هایی هستند که فاصله و اختلاف زیادی با دیگر داده‌ها دارند و تنها یک بار اتفاق افتاده اند و به همین دلیل قابل استناد نیستند. این داده‌ها به عنوان نویز در داده درنظر گرفته می‌شوند و با حذف آنها، روند تغییرات داده تغییر نمی‌کند.

در مجموعه داده اول، داده‌ها مربوط به تراکنش‌های خریدهای انجام شده، شامل کاربر خریدار و کالای خریداری شده‌است. پس از انجام پیش پردازش روی داده ها، مجموعه دادگان منتخب شامل ۴۵۹۱ خرید، ۱۰۶۱ کاربر و ۲۸۳ کالاست. شکل ۳ هیستوگرام ترسیم شده برای تعداد خریدهاست. در این نمودار محور افقی نشان‌دهنده تعداد خرید و محور عمودی نشان‌دهنده تعداد افراد است. به دلیل

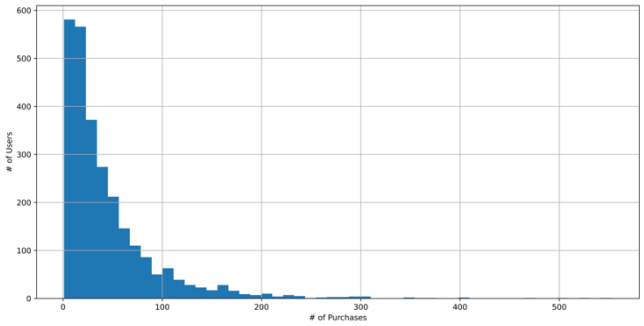
[۱] Feature scaling
[۲] Neighbor-Based Coefficient (NBC)
[۳] Ario (https://ariogames.ir)
[۴] UCI’s Machine Learning Repository
[۵] https://archive.ics.uci.edu/ml/datasets/online+retail
[۶] Outlier

تجمع داده در سمت چپ نمودار و اختلاف مقادیر، به منظور نمایش بهتر از مقیاس لگاریتمی استفاده شده‌است.



شکل ۳. نمودار هیستوگرام تعداد خرید در مجموعه داده Ario

در مجموعه داده دوم نیز داده‌های خرید وجود دارد که مانند بخش قبل، داده‌های مربوط به کاربر خریدار و کالای خریداری‌شده مورد استفاده هستند. پس از پیش‌پردازش و پاکسازی داده‌ها، ۱۲۴۳۸۰ خرید، ۲۶۸۰ کاربر و ۳۵۵۰ کالا در مجموعه قرار دارد. شکل ۴ نمودار هیستوگرام این مجموعه داده در راستای تعداد خرید کاربران است.



شکل ۴. نمودار هیستوگرام تعداد خرید در مجموعه داده UK Retailer

توجه به این نمودارها، داده از نوع توزیع Power Law هستند و بنابراین، طبق مفاهیم تحلیل شبکه اجتماعی، از قاعده انضمام ترجیحی پیروی می‌کند. توضیح آماری این مجموعه داده های منتخب در جدول ۱ آورده شده‌است.

جدول ۱. توصیف آماری مجموعه داده‌ها

معیار	مجموعه داده Ario	مجموعه داده UK Retailer
تعداد	۱۰۶۱	۲۶۸۰
میانگین	۴/۵	۴۵
انحراف از معیار	۶/۵	۵۱
مینیمم	۱	۱
چارک اول	۲	۱۴
میانه	۳	۲۹
چارک سوم	۴	۵۷
ماکزیمم	۱۲۲	۵۵۳

ابتدا از روی این مجموعه داده‌ها، شبکه دوبخشی دادگان ساخته می‌شود. در این شبکه گره‌ها کاربران و کالاها هستند. درصورتی که کاربری کالایی را خریداری کرده باشد، بین آن دو گره، یال تر سیم می گردد. به منظور انجام پیش‌بینی پیوند و پیاده‌سازی روش پیشنهادی، از برنامه نویسی پایتون و کتابخانه‌های آن از جمله pandas، numpy، networkx، sklearn و statistics استفاده شده‌است. خروجی برنامه، یال‌های پیش‌بینی شده توسط هر روش است. نتایج پیاده‌سازی در بخش بعد بررسی شده‌است.

۵ آزمایش و ارزیابی

در بخش ارزیابی، از ۳ شاخص شناخته شده و مرجع در داده‌کاوی تحت عناوین Precision، Recall و F ۱ Score، که دو معیار اول در آن درنظر گرفته شده‌است، استفاده شده است. مجموعه دادگان به صورت تصادفی با نسبت ۰/۷ به ۰/۳، به مجموعه دادگان آموزش و آزمون تقسیم می‌شود. لازم به ذکر است با توجه به اهمیت همبند بودن گراف در تحلیل آن، این تقسیم‌بندی تا زمانی که گراف به دست‌آمده از مجموعه آموزش همبند باشد، ادامه پیدا می‌کند. نتایج به دست آمده از تحلیل و بررسی مجموعه آموزش، با مجموعه آزمون مقایسه و شاخص‌های ارزیابی از روی آن محاسبه می‌شود. به منظور ارزیابی، معیار پیشنهادی با معیارهای تشکیل‌دهنده آن و همچنین ضریب آدامیک-آدار[۲۷]، که میزان شباهت دو گره را با توجه به همسایه‌های مشترک آنها محاسبه می‌کند، مقایسه می‌شود. ضریب آدامیک-آدار به دلیل اهمیت دادن به گره‌های کم‌درجه، در معیار پیشنهادی استفاده نشده‌است. به منظور جلوگیری از جهت‌دار بودن داده‌های انتخابی در هر مجموعه و به دست‌آوردن مجموعه داده عادلانه، این آزمایش ۱۰ بار تکرار می‌گردد و میانگین مقدار شاخص‌های ارزیابی به عنوان مقادیر نهایی درنظر گرفته می‌شوند. جدول ۲ و جدول ۳، پیچیدگی محاسباتی و مقادیر شاخص‌های ارزیابی به درصد را درهر روش به ترتیب برای مجموعه داده‌های اول و دوم نشان می‌دهد.

جدول ۲. نتایج ارزیابی معیارهای پیش‌بینی پیوند روی مجموعه داده Ario

Type	Time Complexity	Precision (%)	Recall (%)	F ۱ Score (%)
JC	$O(n^2)$	۶۲,۸۹	۶۱,۴۳	۶۲,۱۵
AA	$O(n^2)$	**۶۳,۸۶**	۵۹,۰۹	۶۱,۳۸
SC	$O(n^2)$	۶۳,۰۷	۶۱,۸۹	۶۲,۴۸
PA	O(n)	۶۱,۲۵	۵۳,۳۴	۵۷,۰۲
HD	$O(n^2)$	۶۳,۰۵	۶۱,۵۸	۶۲,۳۱

LHN	$O(n^2)$	٦١,٤٦	٥٨,١٨	٦٠,٠٩
NBC	$O(n^2)$	٦٣,٤٧	**٦٢,٩٨**	**٦٣,٢٣**

جدول ۳. نتایج ارزیابی معیارهای پیش‌بینی پیوند روی مجموعه داده UK Retailer

Type	Time Complexity	Precision (%)	Recall (%)	F 1 Score (%)
JC	$O(n^2)$	٩٩,٥٦	٢١,٥٧	٤٣,٢٣
AA	$O(n^2)$	٩٩,٨٠	٢٦,٨٣	٤٢,٣١
SC	$O(n^2)$	٩٦,٦٦	٤٨,٨٠	٦٥,٥٨
PA	$O(n)$	٩٨,٩٦	١٨,٧٧	٣١,٥٩
HD	$O(n^2)$	**٩٩,٨٧**	٢١,٥١	٤٣,١٥
LHN	$O(n^2)$	٩٢,٤٧	٦٣,٩٢	٧٤,٧٨
NBC	$O(n^2)$	٩٣,٦٩	**٦٥,١٣**	**٧٧,٥٤**

در بحث پیش‌بینی پیوند، علاوه بر F 1 Score، Precision نیز شاخصی پراهمیت است. همان‌طور که مشاهده می‌شود، معیار پیشنهادی از نتایج مطلوبی برخوردار است. در هر دو مجموعه داده، معیار پیشنهادی در شاخص‌های F 1 Score و Recall، بالاترین مقادیر میان روش‌ها را دارد. برتری روش پیشنهادی به نسبت دیگر روش‌ها در مجموعه داده UK Retailer که ابعاد بزرگتری دارد، بیشتر مشهود است. تمامی روش‌ها از Precision بالا و مطلوبی برخوردار هستند. اما نکته قابل توجه، کم بودن مقادیر Recall و درنتیجه F 1 Score برای دیگر روش‌هاست که باعث می‌شود روش پیشنهادی روشی مؤثرتر برآورد شود. نتایج ارزیابی معیار پیشنهادی برای مجموعه داده اول Precision ٪۶۳/۴۷، Recall ٪۶۲/۹۸ و F 1 Score ٪۶۳/۲۳ و برای مجموعه داده دوم Precision ٪۹۳/۶۹، Recall ٪۶۵/۱۳ و F 1 Score ٪۷۷/۵۴ به‌دست آمده است. در مجموعه داده اول، ضریب آدامیک-آدار دارای مقدار Precision بالاتری است اما اختلاف دقت این معیار و معیار پیشنهادی، ۰/۳۹ درصد و بسیار ناچیز است. در مجموعه داده دوم، با وجود اینکه دیگر روش‌ها مقادیر Precision بالاتری دارند، اما همانطور که ذکر شد به دلیل عملکرد پایین در دیگر شاخص‌های ارزیابی، این موضوع برتری زیادی را به دنبال ندارد. پیچیدگی محاسباتی این روش‌ها عمدتا از نوع $O(n^2)$ است. با توجه به اینکه پیچیدگی روش پیشنهادی نیز از همین نوع است، در مجموع، معیار پیشنهادی برای استفاده در پیش‌بینی پیوند شبکه دوبخشی عملکرد بهتری به نسبت دیگر روش‌ها نشان داده‌است. نکته مهم دیگر، درنظر گرفتن جنبه‌های مختلف همسایگی و شباهت در این روش است که در این صورت، مسئله از دیدگاهی گسترده‌تر مورد بررسی قرار می‌گیرد.

۶ جمع‌بندی

در این مقاله، روشی مبتنی بر همسایه برای پیش‌بینی پیوند در شبکه دوبخشی ارائه شده‌است. در این روش ابتدا معیارهای شناخته‌شده و کلاسیک پیش‌بینی پیوند مبتنی بر همسایه با تمرکز روی گره‌های پردرجه، با اعمال تغییراتی برای شبکه دوبخشی تعریف می‌شوند. جمع مقادیر نرمال‌شده این معیارها، معیار پیشنهادی را تشکیل می‌دهد. نتایج ارزیابی نشان می‌دهد با وجود پیچیدگی یکسان، روش پیشنهادی از کارایی بهتری به نسبت روش‌های پایه برخوردار است و از فواید معیارهای کلاسیک مختلف بهره می‌برد. به دلیل استفاده از معیارهایی از گروه، هم‌بستگی مفهومی لازم در روش ارائه شده وجود دارد. همچنین استفاده از معیارهای شباهت سنجی و معیارهایی با تمرکز بر گره‌های پردرجه، باعث هم‌افزایی از طریق این دومفهوم و همچنین از طریق معیارهای هرگروه شده‌است که باعث کارایی بیشتر روش پیشنهادی می‌شود. بنابراین هدف پژوهش که دستیابی به یک معیار مبتنی بر هم‌سایه و با کارایی مطلوب در شبکه دوبخشی است، محقق شده است. با این حال، با توجه به در دسترس نبودن مراجع کافی در زمینه مورد مطالعه، قابلیت بهبود روش پیشنهادی وجود دارد. در این روش تأثیر معیارهای پیش‌بینی پیوند با یکدیگر یکسان است. به منظور بهبود کارایی روش، پیشنهاد می‌گردد وزن هر معیار در این راستا بررسی و محاسبه گردد. همچنین برای مطالعه دقیق‌تر، بررسی نتایج روی مجموعه داده‌های دیگر نیز توصیه می‌گردد.

مراجع

[۱] M. Newman, Networks. Oxford university press, ۲۰۱۸.

[۲] A. Alamsyah, "Social network data analytics for market segmentation in indonesian telecommunications industry," ۲۰۱۷, pp. ۱–۵.

[۳] O. Allali, C. Magnien, and M. Latapy, "Link prediction in bipartite graphs using internal links and weighted projection," ۲۰۱۱, pp. ۹۳۶–۹۴۱.

[۴] P. Wang, B. Xu, Y. Wu, and X. Zhou, "Link prediction in social networks: the state-of-the-art," Science China Information Sciences, vol. ۵۸, no. ۱, pp. ۱–۳۸, ۲۰۱۵.

[۵] V. Martínez, F. Berzal, and J.-C. Cubero, "A survey of link prediction in complex networks," ACM computing surveys (CSUR), vol. ۴۹, no. ۴, pp. ۱–۳۳, ۲۰۱۶.

[۶] E. Gündoğan and B. Kaya, "A recommendation method based on link prediction in drug-disease bipartite network," ۲۰۱۷, pp. ۱۲۵–۱۲۸.

[۷] E. Gündoğan and B. Kaya, "A link prediction approach for drug recommendation in

disease-drug bipartite network," ۲۰۱۷, pp. ۱–٤.

[۸] S. Aslan, B. Kaya, and M. Kaya, "Predicting potential links by using strengthened projections in evolving bipartite networks," Physica A: Statistical Mechanics and its Applications, vol. ٥۲٥, pp. ۹۹۸–۱۰۱۱, ۲۰۱۹.

[۹] S. Aslan and B. Kaya, "Time-aware link prediction based on strengthened projection in bipartite networks," Information Sciences, vol. ٥۰٦, pp. ۲۱۷–۲۳۳, ۲۰۲۰.

[۱۰] Y.-J. Chang and H.-Y. Kao, "Link prediction in a bipartite network using wikipedia revision information," ۲۰۱۲, pp. ٥۰–٥٥.

[۱۱] S. Xia, B. Dai, E.-P. Lim, Y. Zhang, and C. Xing, "Link prediction for bipartite social networks: The role of structural holes," ۲۰۱۲, pp. ۱٥۳–۱٥۷.

[۱۲] M. Medo, M. S. Mariani, and L. Lü, "Link prediction in bipartite nested networks," Entropy, vol. ۲۰, no. ۱۰, p. ۷۷۷, ۲۰۱۸.

[۱۳] C. Zhang, E. Chan, and A. Abdulhamid, "Link prediction in bipartite venture capital investment networks," CS ۲۲٤-w report, Stanford, ۲۰۱٥.

[۱٤] W. Wang, X. Chen, P. Jiao, and D. Jin, "Similarity-based regularized latent feature model for link prediction in bipartite networks," Scientific reports, vol. ۷, no. ۱, pp. ۱–۱۲, ۲۰۱۷.

[۱٥] X. Chen, D. Xie, L. Wang, Q. Zhao, Z.-H. You, and H. Liu, "BNPMDA: bipartite network projection for MiRNA–disease association prediction," Bioinformatics, vol. ۳٤, no. ۱۸, pp. ۳۱۷۸–۳۱۸٦, ۲۰۱۸.

[۱٦] D. Zhao, L. Zhang, and W. Zhao, "Genre-based link prediction in bipartite graph for music recommendation," Procedia Computer Science, vol. ۹۱, pp. ۹٥۹–۹٦٥, ۲۰۱٦.

[۱۷] F. Xie, Z. Chen, J. Shang, X. Feng, and J. Li, "A link prediction approach for item recommendation with complex number," Knowledge-Based Systems, vol. ۸۱, pp. ۱٤۸–۱٥۸, ۲۰۱٥.

[۱۸] Y. Cui, L. Zhang, Q. Wang, P. Chen, and C. Xie, "Heterogeneous network linkage-weight based link prediction in bipartite graph for personalized recommendation," Procedia Computer Science, vol. ۹۱, pp. ۹٥۳–۹٥۸, ۲۰۱٦.

[۱۹] Y. Luo, Q. Liu, W. Wu, F. Li, and X. Bo, "Predicting drug side effects based on link prediction in bipartite network," ۲۰۱٤, pp. ۷۲۹–۷۳۳.

[۲۰] L. Zhang, J. Li, Q. Zhang, F. Meng, and W. Teng, "Domain knowledge-based link prediction in customer-product bipartite graph for product recommendation,"

International Journal of Information Technology & Decision Making, vol. ۱۸, no. ۰۱, pp. ۳۱۱–۳۳۸, ۲۰۱۹.

[۲۱] M. Koptelov, A. Zimmermann, B. Crémilleux, and L. Soualmia, "Link prediction via community detection in bipartite multi-layer graphs," ۲۰۲۰, pp. ٤۳۰–٤۳۹.

[۲۲] D. B. Larremore, A. Clauset, and A. Z. Jacobs, "Efficiently inferring community structure in bipartite networks," Physical Review E, vol. ۹۰, no. ۱, p. ۰۱۲۸۰٥, ۲۰۱٤.

[۲۳] G. Salton and J. Michael, "McGill," Introduction to modern information retrieval, vol. ۱, no. ٤,۱, pp. ٤–۱, ۱۹۸٦.

[۲٤] A.-L. Barabâsi, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek, "Evolution of the social network of scientific collaborations," Physica A: Statistical mechanics and its applications, vol. ۳۱۱, no. ۳–٤, pp. ٥۹۰–٦۱٤, ۲۰۰۲.

[۲٥] T. Zhou, L. Lü, and Y.-C. Zhang, "Predicting missing links via local information," The European Physical Journal B, vol. ۷۱, no. ٤, pp. ٦۲۳–٦۳۰, ۲۰۰۹.

[۲٦] E. A. Leicht, P. Holme, and M. E. Newman, "Vertex similarity in networks," Physical Review E, vol. ۷۳, no. ۲, p. ۰۲٦۱۲۰, ۲۰۰٦.

[۲۷] L. A. Adamic and E. Adar, "Friends and neighbors on the web," Social networks, vol. ۲٥, no. ۳, pp. ۲۱۱–۲۳۰, ۲۰۰۳.