استفاده از مدل‌های طبقه‌بندی برای بهینه‌سازی پیش‌بینی لینک در شبکه‌های اجتماعی خودمحور

سهیلا نعمتی [۱]، مهدی صادق زاده [۲*]، مازیار گنجو [۳]

۱: سهیلا نعمتی، گروه مهندسی کامپیوتر، واحد بوشهر، دانشگاه آزاد اسلامی، بوشهر، ایران، s.nemati2012@yahoo.com
۲*: مهدی صادق زاده ، گروه مهندسی کامپیوتر، واحد ماهشهر، دانشگاه آزاد اسلامی، ماهشهر، ایران، sadegh_1999@yahoo.com
۳: مازیار گنجو، گروه مهندسی فناوری اطلاعات، واحد بوشهر، دانشگاه آزاد اسلامی، بوشهر، ایران، ganjoo@gmail.com

تاریخ دریافت: ۱۳۹۸/۹/۲۳ تاریخ پذیرش: ۱۳۹۹/۱/۲۷

چکیده

سیستم‌های پیشنهاد دهنده اجتماعی، نسل جدیدی از این سیستم‌ها می‌باشند که از شبکه اجتماعی به عنوان بستر مدل‌سازی کاربر استفاده می‌کنند تا با استفاده از حجم غنی داده‌های تعاملی، برخی از چالش‌ها را مرتفع نمایند. شبکه‌های آنلاین اجتماعی، دوستان جدید را به کاربران ثبت شده بر مبنای خصوصیات گراف محلی پیشنهاد می‌دهند. هدف اصلی مسئله پیش‌بینی لینک در شبکه‌های اجتماعی، پیشنهاد لیستی از کاربران به یک کاربر خاص می‌باشد که احتمالا در آینده با آنها ارتباط برقرار خواهد کرد. در این تحقیق یک روش پیش‌بینی لینک بر اساس خصوصیات مدل‌های طبقه‌بندی ارائه شده است. در اینجا مسئله پیش‌بینی لینک به یک مسئله طبقه‌بندی با دو کلاس مثبت و منفی تبدیل شده، جائیکه کلاس مثبت نشان دهنده ارتباط و کلاس منفی نشان دهنده عدم ارتباط دو کاربر است. سه طبقه‌بند کلاسیک DT، NN و NB برای کار طبقه‌بندی استفاده شده است. برای ایجاد مجموعه داده از ویژگی‌های اعتبار، خوش‌بینی، تعداد همسایه‌های مشترک، تعداد مسیر با طول‌های متفاوت، تعداد توئییت‌های مشترک، تعداد مسیرهای خود محور داخلی و خارجی بهره گرفته می‌شود. اگر چه شبکه‌های خودمحور همپوشانی زیادی در حلقه‌ها ندارند، اما آزمایش‌ها نشان می‌دهد که در نظر گرفتن اطلاعات مسیرهای خودمحور به طور قابل توجهی عملکرد پیش‌بینی را بهبود می‌بخشد. طبقه‌بندی DT بهترین عملکرد را با دقت متوسط ٪99.85 به ثبت رسانیده است.

واژه های کلیدی: شبکه‌های اجتماعی خودمحور، معیار شباهت، طبقه‌بندی داده‌ها، استخراج ویژگی.

۱- مقدمه

شبکه‌های اجتماعی، نسل جدیدی از پایگاه‌هایی می‌باشند که این روزها در کانون توجه کاربران شبکه جهانی اینترنت قرار دارند. این گونه پایگاه‌ها بر مبنای تشکیلات آنلاین فعالیت کرده و هر کدام دسته‌ای از کاربران اینترنتی با ویژگی خاص را گرد هم آورده‌اند. شبکه‌های اجتماعی را گونه‌ای از رسانه‌های اجتماعی می‌دانند که امکان دستیابی به شکل جدیدی از برقراری ارتباط و به اشتراک گذاری محتوا در اینترنت را به وجود آورده‌اند [۱]. با گسترش فضای مجازی و شبکه‌های اجتماعی و فراگیر شدن آن در بین جوامع مختلف بحث مربوط به تحلیل این شبکه‌ها در مطالعات و تحقیقات بسیاری مورد توجه قرار گرفته است [۲].

شبکه اجتماعی، عموماً ساختاری از گره‌های فردی یا سازمانی می‌باشند. تحلیل شبکه‌های اجتماعی، روابط اجتماعی را با اصطلاحات رأس و یال می‌نگرد [۳]. رأس‌ها بازیگرهای فردی درون شبکه‌ها می‌باشند و یال‌ها روابط میان این بازیگران هستند. بسیاری از یال‌ها می‌توانند میان رأس‌ها وجود داشته باشند. در رابطه با این لینک‌ها و ارتباطات آنها، مساله پیش‌بینی لینک که امری مهم برای تحلیل شبکه‌های اجتماعی می‌باشد، اهمیت پیدا کرده است. این مساله به معنی پیش‌بینی احتمال برقراری یک ارتباط بین دو رأس با دانستن این موضوع که در حال حاضر بین این دو رأس ارتباطی وجود ندارد، می‌باشد [۴]. با افزایش محبوبیت وب، داده‌ها در شبکه‌های بزرگ میتوانند برای مطالعه تکامل شبکه‌های اجتماعی و جریان اطلاعات در سطح بسیار کوچک بازیابی شوند. تکامل چنین داده‌هایی از شبکه را میتوان در یک ساختار واحد (گراف تحول زمان) نشان داد، زیرا داده‌های این شبکه‌ها با گذشت زمان از بین نمی‌روند بلکه رشد

می‌کنند. با توجه به تصویر مختصری از گراف، مسئله پیش‌بینی لینک میتواند داده‌های شبکه‌های بزرگ را با گذشت زمان تجزیه و تحلیل کند. پیش‌بینی لینک یک چالش مهم در تحلیل شبکه‌های اجتماعی است که کاربردهایی در حوزه‌های دیگر نظیر بازیابی اطلاعات، بایوانفورماتیک و تجارت الکترونیک نیز دارد. علاوه بر این، امروزه با توجه به فعالیت همزمان کاربران در چندین شبکه این ضرورت وجود دارد که برای تحلیل دقیق‌تر شبکه‌های اجتماعی از داده‌های چندین گراف شبکه استفاده شود. بنابراین، مسئله پیش‌بینی لینک در شبکه‌های چندلایه مطرح می‌شود. پیش‌بینی لینک در شبکه‌های تک لایه با مواردی مطابقت دارد که فقط اطلاعات داخل لایه مورد نظر در نظر گرفته شود و این مسئله در شبکه‌های چند لایه با مواردی مطابقت دارد که علاوه بر اطلاعات داخل لایه از شبکه هدف، اطلاعات بین لایه‌ها در شبکه‌های دیگر نیز در نظر گرفته شوند.

اغلب پیش‌بینی لینک با استفاده از سیستم‌های توصیه‌گر در شبکه‌های اجتماعی مورد بررسی قرار می‌گیرد. سیستم‌های توصیه‌گر، شاخه‌ای از سیستم‌های بازیابی و تطبیق اطلاعات می‌باشند که با شناسایی علاقمندی‌ها و نیازمندی‌های کاربر، به آنها در دستیابی به اطلاعات یا خدمات مورد نظر در میان حجم انبوهی از انتخاب کمک می‌کنند. عدم وجود داده کافی و پراکندگی داده‌ها، از جمله چالش‌های پردازش توصیه برای سیستم‌های توصیه‌گر است. شبکه‌های آنلاین اجتماعی، دوستان جدید را به کاربران ثبت شده بر مبنای خصوصیات گراف محلی (تعداد دوستان مشترک میان دو کاربر) پیشنهاد می‌دهد. هر چند شبکه‌های آنلاین اجتماعی از تمام ظرفیت‌های مسیرشبکه استفاده نمی‌کنند. آنها تنها بر روی مسیرهای با حداکثر طول دو میان یک کاربر و دوست منتخبش تمرکز دارند. از سوی دیگر، رویکردهای سراسری وجود دارند که تمامی ساختارهای مسیری در یک شبکه را که از نظر محاسباتی برای شبکه‌های اجتماعی بزرگ بسیار سخت است، شناسایی می‌کنند [۵، ۶].

در این تحقیق ابتدا به بررسی و مقایسه جدیدترین روش‌های ارائه شده در چند سال اخیر جهت پیشنهاد دوست در شبکه‌های اجتماعی می‌پردازیم و سپس با ترکیب برخی از این روش‌ها یک سیستم پیش‌بینی لینک جدید برای شبکه‌های اجتماعی چندگانه ارائه می‌دهیم. استفاده از ویژگی‌های ساختاری و توپولوژی گراف در جهت تبدیل مسئله پیش‌بینی لینک به یک مسئله طبقه‌بندی از اهداف اصلی این تحقیق می‌باشد. برای اینکار از نمونه‌های واقعی ارائه شده توسط شبکه‌های اجتماعی معروف نظیر توئیتر استفاده می‌کنیم. تاکنون روش‌های زیادی برای حل مسئله پیش‌بینی لینک جدید در شبکه‌های اجتماعی چندگانه پیشنهادی شده است که اغلب مبتنی بر معیارهای شباهت توپولوژی هستند. در این میان برخی روش‌ها نیز از الگوریتم‌های طبقه‌بندی برای مدلسازی مسئله استفاده می‌کنند. کارایی این روش‌ها مبتنی بر استخراج ویژگی‌ها و نحوه ساخت مجموعه داده هستند. بطور کلی نوآوری روش پیشنهادی و وجه تمایز آن با روش‌های مشابه، استخراج ویژگی‌های وابسته به شبکه‌های خود-محور است که تاکنون برای اینکار استفاده نشده است.

در ادامه این تحقیق به بررسی برخی از کارهای انجام شده در بخش ۲ می‌پردازیم. در بخش ۳ الگوریتم پیشنهادی مبتنی بر تبدیل مسئله به یک مسئله طبقه‌بندی ارائه می‌شود. نتایج ارزیابی روش پیشنهادی و بحث در مورد آن در بخش ۴ آورده شده و در نهایت نتیجه‌گیری و پیشنهادها در بخش ۵ ذکر شده است.

۲- کارهای انجام شده

با رشد سریع شبکه های اجتماعی، کاربران این شبکه با یک مقدار زیادی از اطلاعات مواجه هستند. توصیه دوست در یکی از محبوب‌ترین شبکه‌های اجتماعی آنلاین به کاربران برای یافتن دوستان جدید و گسترش چرخه اجتماعی کمک می‌کند. در زمینه پیش بینی لینک در شبکه های اجتماعی، تحقیقات متعددی صورت گرفته است. در این بخش به تحلیل و بررسی برخی از جدیدترین تحقیق های مرتبط در زمینه می پردازیم. ناول و کلینبرگ [۷] اولین مدل پیش‌بینی لینک که صراحتاً در شبکه‌های اجتماعی کاربرد داشت را ارائه کردند [۷]. روش پیش‌بینی بر اساس تشابه بین دو گره است که احتمال دارد در آینده با هم دوست شوند. آنها گره‌ها را بر اساس میزان امتیاز شباهت‌هایشان رتبه‌بندی کردند. بعد از آنها محمد [۸] این رویکرد را با دو روش دیگر گسترش داد [۸].

جلیلی و همکاران [۹]، مسئله پیش‌بینی لینک در شبکه‌های اجتماعی آنلاین چندگانه را مبتنی بر روش‌های طبقه‌بندی مورد مطالعه قرار دادند [۹]. آنها یک الگوریتم مبتنی بر متا-مسیر برای پیش‌بینی لینک با توجه به دو مدل طبقه‌بندی توسعه دادند. سیو و همکاران[۱۰]، یک روش پیش‌بینی لینک محدود در شبکه‌های اجتماعی بسیار بزرگ ارائه دادند [۱۰]. در این تحقیق با در نظر گرفتن

کران پایین ارزش شباهت بین هر جفت گره و همچنین با تمرکز بر روی معیار CN (همسایگان مشترک)، این مشکل را تا حدی برطرف کرده‌اند. در تحقیق دیگری پروازه و همکاران [۱۱]، از نظریه گراف و مشخصات کاربران برای دوست‌یابی در شبکه های اجتماعی استفاده کردند [۱۱]. در مرحله اول سیستم پیشنهادی، تمام کاربران با توجه به شباهت ساختاری و پروفایل و با استفاده از الگوریتم K_Medoids خوشه‌بندی می‌شوند. در مرحله دوم، الگوریتم FriendLink به کاربران هر خوشه اعمال شده و میزان تشابه بین کاربران محاسبه می‌شود.

لیو و همکاران [۱۲]، یک روش پیش‌بینی لینک مبتنی بر شبکه‌های باور عمیق را در شبکه‌های اجتماعی تحت امضاء ارائه دادند [۱۲]. آنها از اصول پنهان رفتارهای اعضای اجتماعی که به ندرت انجام شده، به منظور پیش‌بینی لینک بهره گرفتند. لایشرام و همکاران [۱۳]، روش‌های موجود در پیش‌بینی لینک را گسترش داده و یک الگوریتم یادگیری با ناظر را برای پیش‌بینی لینک در شبکه‌هایی با لینک‌های غیر دائم ارائه کردند [۱۳]. در این تحقیق مسئله پیش‌بینی لینک به صورت یک مسئله طبقه‌بندی مطرح شده است. در تحقیق دیگری نادری‌پور و همکاران [۱۴]، یک روش محاسبات «دانه دانه» و یک منطق فازی دو مدله را برای پیش‌بینی لینک‌ها و رابطه بین دو گره ارائه دادند [۱۴]. این روش روی شبکه‌های همکاری تست شده است.

سیزپانسکی و همکاران [۱۵] یک معیار جدید از شباهت بین گره‌ها بر اساس شاخص متقابل تئوری بازی ارائه دادند [۱۵]. این معیار بر دو مسئله پیش‌بینی لینک و تشخیص جامعه در شبکه‌های اجتماعی تاکید دارد و در زمان چند جمله‌ای محاسبه می‌شود. ژائو و همکاران [۱۶]، پیش‌بینی لینک‌ها و وزن‌ها در شبکه‌ها را با مسیرهای قابل اطمینان ارائه نمودند [۱۶]. هدف از پیش‌بینی لینک کشف لینک‌های گم شده و یا پیش‌بینی ظهور روابط در آینده با توجه به ساختار شبکه جاری می‌باشد. در تحقیق دیگری هان و خو [۱۷] با توجه به ماهیت شبکه اجتماعی میکروبلاگ، یک سیستم پیش‌بینی لینک با ترکیب ویژگی‌های متعدد توسط مطرح کردند[۱۷].

ژانگ و همکاران [۱۸]، مسئله فراتر از پیش‌بینی لینک در فضای مجاور را مطرح ساختند [۱۸]. در این تحقیق، مسئله پیش‌بینی فرالینک در فراشبکه مطرح شده که هدف پیش‌بینی لینک برای گره‌های چندگانه است. هو و همکاران [۱۹]، پیش‌بینی لینک با نفوذ اجتماعی شخصی را پیشنهاد دادند [۱۹]. در این تحقیق با توجه به نشانه‌های زمانی هر کاربر و فعالیت‌های اجتماعی آنها، از آنتروپی برای اندازه‌گیری کاهش عدم قطعیت همسایگان استفاده می‌شود. تأثیرات اجتماعی آماری سپس به یک مدل پیش‌بینی لینک مبتنی بر گراف برای انجام یادگیری مشترک اضافه می‌شود. در تحقیق دیگری پی و همکاران [۲۰]، پیش بینی پیوند در شبکه های پیچیده براساس یک شاخص تخصیص اطلاعات را ارائه دادند [۲۰]. آنها برای پیش‌بینی لینک یک شاخص جدید که مبتنی بر همسایگی با فرایند تخصیص اطلاعات مجازی است، طراحی کردند.

لسکواک و مکاولی [۲۱]، یک روش خوشه‌بندی در جهت یادگیری در کشف حلقه‌های اجتماعی برای شبکه‌های خودمحور توسعه دادند [۲۱]. هدف این روش پیش‌بینی مجموعه نهایی حلقه‌های کاربر است. شانگ و همکاران [۲۲] نقش لینک‌های مستقیم برای پیش‌بینی لینک در شبکه‌های در حال تحول را مطرح ساختند [۲۲]. در اینجا زمان به عنوان پارامتری برای بررسی دقت پیش‌بینی لینک معرفی شده است. شارما و سینگ [۲۳]، یک روش کارآمد برای پیش‌بینی لینک در شبکه‌های چندگانه وزن‌دار ارائه دادند [۲۳]. در این تحقیق از درجه لینک‌ها با استفاده از شاخص‌های شباهت به منظور پیش‌بینی وزن‌ها بهره گرفته می‌شود. آناگنستوپولوس و همکارن [۲۴]، تشخیص جامعه بر روی گراف‌های در حال تکامل را مطرح ساختند که می‌توانست دوستان مشابه را در دسته‌های یکسانی قرار دهد [۲۴]. پاپادیمیتریو و همکاران [۲۵]، یک الگوریتم مبتنی بر پیش‌بینی سریع و دقیق ارتباطات در شبکه‌های اجتماعی مطرح ساختند [۲۵]. در اینجا از الگوریتم FriendLink برای رسیدن به مقیاس‌پذیری استفاده شده است. Friendlink می‌تواند ماتریس تشابه میان دو گره گراف را پیدا کند و دوستان را بر اساس اهمیت پیشنهاد دهد. بای و همکاران [۲۶]، خوشه‌بندی سریع گراف با یک مدل توصیف جدید برای تشخیص جامعه را مطرح ساختند [۲۶]. توصیف و کشف موثر جوامع در یک شبکه، مفهوم تحقیقاتی مهمی برای دسته‌بندی گراف و پیش‌بینی لینک است. آقابزرگی و خیامامشی [۲۷]، یک معیار شباهت جدید برای پیش‌بینی لینک در شبکه‌های اجتماعی ارائه دادند [۲۷]. معیار شباهت پیشنهاد شده از توزیع رأس‌ها در موتیف‌های شبکه ساخته شده است. موتیف‌ها بلوک‌های کوچک ساختمانی از شبکه‌های اطلاعاتی هستند. تأثیر این موتیف‌ها، تفاوت اصلی بین معیار شباهت پیشنهادی و دیگر معیارهای شباهت بر پایه همسایگی است.

۳- الگوریتم پیشنهادی

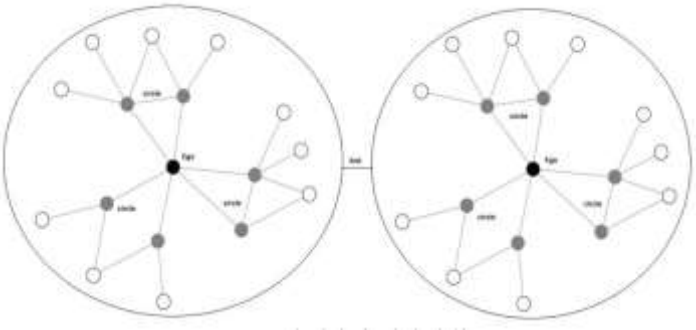
پیش‌بینی یکی از جنبه‌های جذاب در داده‌کاوی است و اخیراً توجه بسیاری از محققان را به عنوان یک روش مؤثر برای تجزیه و تحلیل شبکه اجتماعی خود جلب کرده است. هر چند این مسئله به صورت گسترده‌ای مورد مطالعه و بررسی قرار گرفته است؛ با این

حال، مشکل چگونگی ترکیب بهینه و مؤثر اطلاعات توپولوژیکی حاصل از ساختار شبکه با داده‌های توصیفی فراوان مربوط به گره و یال، تا حد زیادی پا برجا است. با محبوبیت شبکه اجتماعی، پیش‌بینی لینک در شبکه‌های اجتماعی چندگانه یک چالش مورد توجه در میان محققان شده است. در بیشتر پژوهش‌های انجام شده در این زمینه، ارتباطات درون لایه‌ای که نشان دهنده قدرت ارتباط است در نظر گرفته نشده است، درحالیکه این ارتباطات حاوی اطلاعات مفیدی در این راستا می‌باشد. همچنین میتوان از اطلاعات ساختاری دیگری مانند اطلاعات شبکه‌های خود-محور نیز برای بهبود کارایی پیش‌بینی لینک استفاده نمود.

در حال حاضر، تقریباً همه الگوریتم‌های پیش‌بینی لینک برای شبکه‌های چندگانه تنها روی اطلاعات ساختار توپولوژیکی تمرکز دارند و اطلاعات درون لایه‌ای مربوط به شبکه‌های خود-محور را نادیده می‌گیرند. بنابراین، این روش‌ها از چالش کمبود اطلاعات توپولوژیکی گراف شبکه رنج می‌برند. جنبه علمی و جدید بودن این تحقیق در نظر گرفتن همزمان اطلاعات درون لایه‌ای شبکه‌های خودمحور و همچنین اطلاعات ساختار توپولوژیکی برای بهبود حل مسئله پیش‌بینی لینک در شبکه‌های اجتماعی است. علاوه بر این، به منظور افزایش دقت و سرعت در پروسه پیش‌بینی لینک از مدل‌های طبقه‌بندی برای یادگیری بهره گرفته شده است.

۳-۱- ساختار شبکه‌های اجتماعی خودمحور

در این تحقیق از ساختار شبکه‌های خودمحور استفاده شده است که ارتباط‌های بین کاربران آن جهت‌دار می‌باشد. شبکه‌های خودمحور شبکه‌هایی هستند که در آن‌ها روابط فرد و دوستانش سنجیده و به تصویر کشیده می‌شود [۲۸]. در تجسم گراف چنین شبکه‌هایی، کاربر خود (Ego) معمولا در مرکز گراف قرار می‌گیرد. شکل ۱ ساختار یک شبکه Ego را به تصویر می‌کشد.

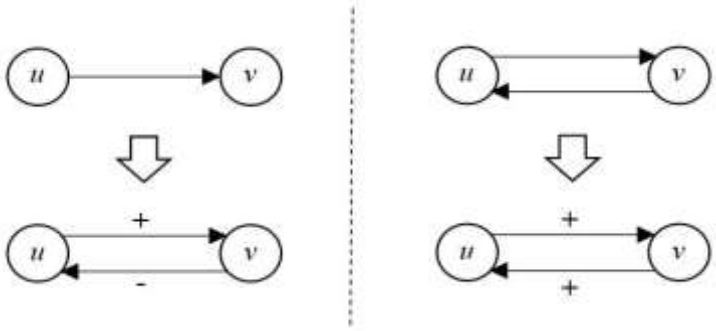


شکل ۱ : ساختار شبکه‌های اجتماعی خودمحور

هر شبکه Ego حول یک کاربر تشکیل شده است که کاربر Ego نامیده می‌شود(گره‌های مشکی در شکل). در هر شبکه Ego تعدادی «حلقه» وجود دارد (گره‌های خاکستری نمایانگر حلقه‌ها در شکل هستند). در هر حلقه بین چند کاربر لینک وجود دارد و با یکدیگر در ارتباط هستند.

۳-۲- ایجاد شبکه علامت‌دار

در این تحقیق مسئله پیش‌بینی لینک را به یک مسئله طبقه‌بندی با دو کلاس مثبت و منفی تبدیل می‌کنیم، جائیکه کلاس مثبت نشان دهنده ارتباط و کلاس منفی نشان دهنده عدم ارتباط دو کاربر است. بنابراین، ساختار شبکه را با تخصیص نشانه‌های مناسب به لینک‌ها اصلاح می‌کنیم. به هر لینک در شبکه یک نشانه مثبت یا منفی اعمال می‌شود. فرض کنید u و v دو گره در یک شبکه اجتماعی باشد. اگر این دو گره یکدیگر را دنبال کنند، لینک بین آنها دوطرفه است، یعنی دو لینک جهت‌دار بین آنها وجود دارد (از u به v و برعکس). در چنین مواردی، علامت مثبت در هر دو جهت قرار داده می‌شود. در مواردی که فقط یکی از گره‌ها دیگری را دنبال می‌کند، علامت مثبت در جهت دنبال‌کننده و علامت منفی در جهت مخالف قرار داده می‌شود. شکل ۲ چگونگی تخصیص علامت به لینک‌ها را نشان می‌دهد.



شکل ۲ : تخصیص نشانه‌های مناسب به لینک‌ها در شبکه‌های اجتماعی

۳-۳- استخراج ویژگی‌های موثر از گراف

برای ایجاد یک مجموعه داده نیاز به استخراج ویژگی‌های موثری از گراف بین کاربران می‌باشد. در این تحقیق از ۹ ویژگی برای ایجاد یک نمونه بین دو گره u و v بهره گرفته شده است.

۳-۳-۱- ویژگی اعتبار برای گره u

اعتبار به محبوبیت گره در شبکه اجتماعی اشاره دارد. این روش مقادیر بالاتری را برای یک گره محاسبه می‌کند، به همین دلیل این گره بیشتر برای دیگران مورد پذیرش است و با احتمال بیشتری آن را دنبال می‌کنند. فرض کنید $d_{in}^{+}(u)$ و $d_{in}^{-}(u)$ به ترتیب تعداد لینک‌های ورودی مثبت و منفی به u باشد. ویژگی اعتبار نرمال شده [RP] به صورت زیر محاسبه می‌شود(رابطه‌ی ۱) [۹].

$$RP(u) = 1 + \frac{d_{in}^{+}(u) - d_{in}^{-}(u)}{d_{in}^{+}(u) + d_{in}^{-}(u)} \quad (۱)$$

با توجه به اینکه در شبکه‌های اجتماعی و فرایند علامت گذاری لینک‌ها، تعداد کاربرانی که بدون ارتباط هستند بسیار زیاد است، لذا یک واحد به مقدار RP به منظور جلوگیری از منفی شدن ترم اضافه می‌کنیم.

۳-۳-۲- ویژگی اعتبار برای گره v

مقدار *RP(v)* نیز مشابه *RP(u)* محاسبه می‌شود.

۳-۳-۳- ویژگی خوش‌بینی برای گره u

گره‌هایی با مقدار خوش‌بینی بالاتر، احتمالا به گره‌هایی با علامت مثبت دیگری سوق پیدا می‌کنند. به این دلیل که خود آنها نیز احتمالا به دنبال ارتباط با سایر گره‌ها هستند. فرض کنید که $d_{out}^{+}(u)$ و $d_{out}^{-}(u)$ به ترتیب تعداد لینک‌ها با خروجی مثبت و منفی از u باشد. ویژگی خوش‌بینی نرمال شده (OP) از u به صورت زیر محاسبه می‌شود (رابطه ۲) [۹].

$$OP(u) = 1 + \frac{d_{out}^{+}(u) - d_{out}^{-}(u)}{d_{out}^{+}(u) + d_{out}^{-}(u)} \quad (۲)$$

۳-۳-۴- ویژگی خوش‌بینی برای گره v

مقدار OP(v) نیز مشابه OP(u) محاسبه می‌شود.

۳-۳-۵- ویژگی تعداد همسایه‌های مشترک برای دو کاربر u و v

این ویژگی تعداد همسایه‌های مشترک بین u و v را اندازه‌گیری می‌کند. بنابراین CN(u,v) تعداد گره‌هایی را نشان می‌دهد که بین هر دو گره u و v لینک دارند و از طریق رابطه ۳ بدست می‌آید [۱۰].

$$CN(u,v) = |\Gamma(u) \cap \Gamma(v)| \quad (۳)$$

در این رابطه و $\Gamma(v)$ به ترتیب همسایگان گره u و v را نشان می‌دهد.

۳-۳-۶- ویژگی تعداد مسیرهای با طول L بین دو کاربر u و v

این معیار با توجه به تعداد و طول مسیرهای میان دو کاربر میزان شباهت بین آنها را تخمین می‌زند. معیار طول مسیرهای متفاوت (PL) به شاخص کاتز مشهور است و به صورت رابطه‌ی ۴ تعریف می‌شود [۱۶].

$$PL(u,v) = \sum_{l=1}^{L} \beta^{l} \cdot \left|paths_{u,v}^{l}\right| \quad (۴)$$

در این رابطه مقدار $|paths_{u,v}^{l}|$ تعداد مسیرهای به طول l بین دو کاربر u و v است. برای کاهش اثر مسیرهای طولانی در تعیین امتیاز یک لینک، از ضریب ثابت β استفاده شده است.

۳-۳-۷- ویژگی تعداد توئییت‌های مشترک بین دو کاربر u و v

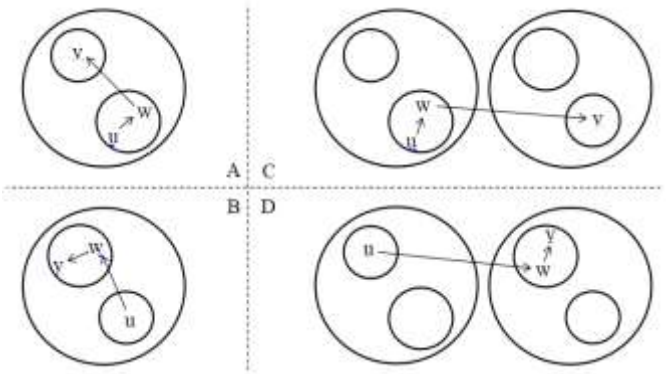
هر چه دو کاربر توئیت‌های استفاده شده مشابهی داشته باشند، احتمال اینکه در آینده بین آنها ارتباط ایجاد شود بیشتر می‌باشد. ویژگی تعداد توئییت‌های مشترک (CT) برای دو کاربر u و v به صورت رابطه‌ی ۵ تعریف می‌شود.

$$CT(u,v) = \sum_{t=1}^{T} feat_{u,t} \cap feat_{v,t} \qquad (۵)$$

در این رابطه T تعداد کل توئیت‌ها، $feat_{u,t}$ و $feat_{v,t}$ به ترتیب t-مین توئیت برای کاربران u و v است.

۳-۳-۸- تعداد مسیرهای خود محور داخلی بین دو کاربر u و v

این ویژگی مبتنی بر استراتژی ابتکاری Ego-Path است و براساس شبکه‌های خودمحور در این تحقیق ارائه شده است. در یک شبکه اجتماعی مبتنی بر Ego، دو گره می‌توانند با مسیرهایی که از حلقه‌های مختلف شبکه و یا از چندین شبکه Ego عبور می‌کنند، ارتباط داشته باشند. برای تعریف این ویژگی ابتدا باید انواع Ego-Pathهای بین دو گره، با توجه به یک طول مسیر از پیش تعیین شده براساس عبور از طرح شبکه پیشنهادی استخراج شوند. در این تحقیق ویژگی تعداد مسیرهای خود محور داخلی (IEP) بین دو گره u و v به صورت $u \xrightarrow{follow} w \xrightarrow{friends} v$ نشان داده می‌شود (u به دنبال w است و w نیز با v دوست است)، جائیکه w گره‌ای است که باعث ارتباط بین دو گره u و v می‌شود. در ویژگی IEP گره w باید حداقل با یکی از دو گره u و v در یک حلقه باشد (ارتباط درون حلقه‌ای). تعداد گره‌های w با این شرایط مقدار این ویژگی را مشخص می‌کنند. شکل ۳ مثالی از IEP بین دو گره u و v از طریق w با طول مسیر ۲ را نشان می‌دهد.

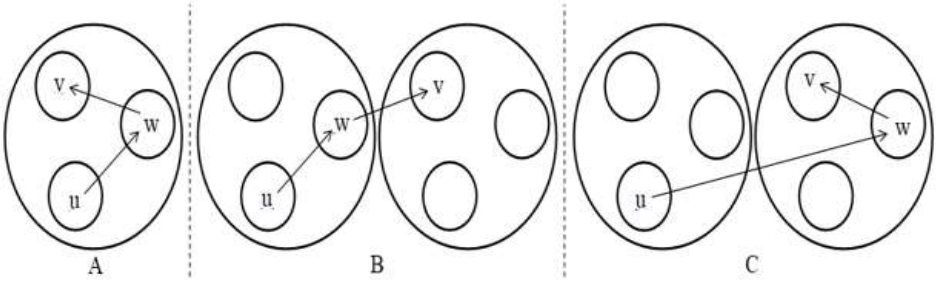


شکل ۳ : مثالی از ویژگی تعداد مسیرهای خود محور داخلی (IEP)

در این شکل دایره‌های بزرگ نقش شبکه‌های Ego و دایره‌های داخلی (کوچک) نقش حلقه‌ها را دارند. چهار حالت مختلف برای این ویژگی وجود دارد. حالتی که گره w و u در یک حلقه و گره v در حلقه دیگری در یک شبکه Ego قرار داشته باشند (حالت A). حالتی که گره w و v در یک حلقه و گره u در حلقه دیگری در یک شبکه Ego قرار داشته باشند (حالت B). حالتی که گره w و u در یک حلقه و گره v در حلقه دیگری در دو شبکه Ego مختلف قرار داشته باشند (حالت C). و درنهایت حالتی که گره w و v در یک حلقه و گره u در حلقه دیگری در دو شبکه Ego قرار داشته باشند (حالت D). ویژگی IEP به ازای تعداد گره‌های w یافت شده برای چهار حالت محاسبه می‌شود.

۳-۳-۹- تعداد مسیرهای خود محور خارجی بین دو کاربر u و v

در این تحقیق ویژگی تعداد مسیرهای خود محور خارجی (IEP) بین دو گره u و v به صورت $u \xrightarrow{follow} w \xrightarrow{friends} v$ نشان داده می‌شود، جائیکه w گره‌ای است که باعث ارتباط بین دو گره u و v می‌شود. در ویژگی OEP گره w باید حداقل با یکی از دو گره u و v در دو حلقه مختلف باشد (ارتباط برون حلقه‌ای). تعداد گره‌های w با این شرایط مقدار این ویژگی را مشخص می‌کنند. شکل ۴ مثالی از OEP بین دو گره u و v از طریق w با طول مسیر ۲ را نشان می‌دهد.



شکل ۴ : مثالی از ویژگی تعداد مسیرهای خود محور خارجی (OEP)

سه حالت مختلف برای این ویژگی وجود دارد. حالتی که سه گره w، u و v در حلقه‌های مختلفی از یک شبکه Ego قرار داشته باشند (حالت A). حالتی که گره‌های w و u در دو حلقه مختلف از یک شبکه Ego و گره v در یک حلقه از شبکه Ego دیگری قرار داشته باشند (حالت B). و درنهایت حالتی که گره u در یک حلقه از یک شبکه Ego و گره‌های w و v در دو حلقه مختلف از شبکه Ego دیگری قرار داشته باشند (حالت C).

۳-۴- ایجاد مدل طبقه‌بندی

برای ایجاد مجموعه داده، از ۹ ویژگی استخراج شده استفاده می‌کنیم. این ویژگی‌ها برای هر جفت کاربر، یک نمونه از مجموعه داده را نشان می‌دهد. کلاس هر نمونه، وضعیت ارتباط دو کاربر (u,v) را نشان می‌دهد که دو حالته (مثبت و منفی) می‌باشد. کلاس مثبت (عدد 1) وجود ارتباط و کلاس منفی (عدد 0) عدم وجود ارتباط را بین دو کاربر نشان می‌دهد. با تبدیل مسئله به یک مسئله طبقه‌بندی با دو کلاس مثبت و منفی، از یک مدل طبقه‌بندی برای آموزش داده‌ها استفاده می‌کنیم. در این تحقیق از سه طبقه‌بند کلاسیک DT، NN و NB به منظور مدل کردن داده‌های استخراج شده استفاده می‌شود. هدف ما بررسی نتایج هر یک از این طبقه‌بندها در تشخیص کلاس نمونه‌های ورودی جدید می‌باشد.

۳-۵- پیشنهاد لینک

در این تحقیق با توجه به استفاده از مدل طبقه‌بندی برای نشان دادن ارتباط بین دو کاربر، از مقدار کلاس پیش‌بینی شده توسط طبقه‌بند به عنوان شباهت بهره می‌گیریم. برای محاسبه شباهت بین دو کاربر آزمایش و آموزشی، ابتدا ویژگی‌های مورد نیاز مدل طبقه‌بندی را استخراج شده، سپس کلاس (0 یا 1) نمونه با استفاده از مدل طبقه‌بند تشخیص داده می‌شود. با توجه به در نظر گرفتن top-k کاربر با بالاترین میزان شباهت جهت پیشنهاد، اگر تعداد کاربرانی با کلاس پیش‌بینی شده 1 کمتر از مقدار top-k باشد، این مقدار با توجه به تعداد پیش‌بینی‌ها کلاس 1 بروز می‌شود. top-k تعداد کاربران با بالاترین شباهت را جهت پیشنهاد نشان می‌دهد.

۴- نتایج و بحث

در این بخش، نتایجی که بدست آمده از روش پیشنهادی را بررسی می‌کنیم و نشان می‌دهیم که این روش نتایج مطلوب‌تری را نسبت به سایر روش‌های مشابه گزارش می‌دهد. نتایج حاصل از شبیه‌سازی روش پیشنهادی با عنوان «Ego-Path» در تمام آزمایش‌ها نشان داده شده است. در این تحقیق از الگوریتم‌های «پیوند دوست»، «کاتز» و «پیش‌بینی لینک مبتنی بر متا-مسیر» با یک سری داده‌های واقعی از شبکه‌های اجتماعی توئیتر و فیسبوک جهت مقایسه و ارزیابی عملکرد روش پیشنهادی بهره گرفته شده است. در این تحقیق برای ارزیابی روش پیشنهادی و سایر روش‌های مورد مقایسه از اعتبارسنجی 10-Fold در تمام آزمایش‌ها استفاده شده است. کاربران شبکه اجتماعی در هر مرحله از اعتبارسنجی به دو بخش داده‌های آموزشی (E^T) و داده‌های آزمایشی (E^P) تقسیم می‌شوند. مشخص

است که $E = E^T \cup E^P$ و $E^T \cap E^P = \emptyset$ است، جائیکه E کل مجموعه داده را نشان می‌دهد. برای هر کاربر آزمایش که حداقل یک لینک در (E^P) دارد، پیشنهاد لینک بر اساس دوستان او در(E^T) ارائه می‌شود.

روش ارائه شده در این تحقیق بر مبنای شبکه‌های خود محور (Ego) می‌باشد، بنابراین برای انجام آزمایش‌ها و ارزیابی عملکرد روش پیشنهادی از دو مجموعه داده خودمحور توئیتر و فیسبوک استفاده می‌شود که از وب‌سایت SNAP دردسترس می‌باشند. هر یک از این مجموعه داده‌ها از تعدادی شبکه Ego تشکیل شده است. هر شبکه یک کاربر Ego با تعدادی «حلقه» دارد (کاربر Ego ممکن است در چندین حلقه وجود داشته باشد). هر حلقه لیستی از کاربرانی را نشان می‌دهد که با همدیگر ارتباط دارند. با توجه به حجم بالای تعداد گره‌ها و لینک‌ها در مجموعه داده‌های دو شبکه اجتماعی توئیتر و فیسبوک، در این تحقیق به منظور ارزیابی عملکرد روش پیشنهادی بخشی از اطلاعات کاربران مطابق با جدول ۱ را در نظر می‌گیریم.

جدول ۱ : مجموعه داده‌های استفاده شده در شبیه‌سازی

	نسخه اصلی			نسخه استفاده شده		
مجموعه داده	تعداد شبکه‌های Ego	تعداد گره‌ها	تعداد لینک‌ها	تعداد شبکه‌های Ego	تعداد گره‌ها	تعداد لینک‌ها
توئیتر	۹۷۳	۸۱۳۰۶	۱۷۶۸۱۴۹	۲۰	۲۱۰۱	۴۵۵۶۹
فیسبوک	۱۰	۴۰۳۹	۸۸۲۳۴	۳	۲۱۶۷	۴۵۴۲۳

پارامترهای استفاده شده در شبیه‌سازی به منظور مقایسه دقیق‌تر بین روش‌ها یکسان در نظر گرفته می‌شود. مقادیر این پارامترها به صورت $th = 0.01$، $\beta = 0.1$ و $top - k = 10, 20$ تنظیم شده است. برای اندازه‌گیری معیارها در هر کاربر هدف (آزمایش)، میزان شباهت را با تمام کاربران آموزشی محاسبه کرده و لیستی از top-k کاربر با بالاترین شباهت را به کاربر هدف پیشنهاد می‌دهیم.

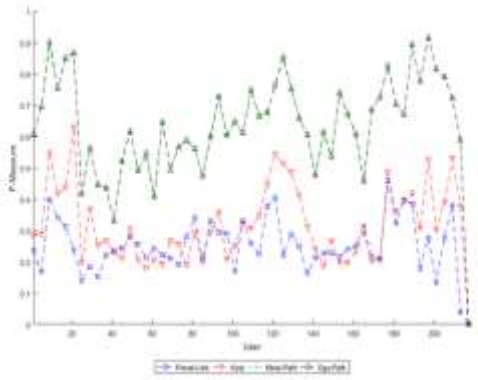
در روش‌های متا-مسیر و Ego-Path جهت طبقه‌بندی مجموعه داده استخراج شده از سه طبقه‌بند DT، NN و NB استفاده می‌شود. شکل ۵ نتایج معیار اندازه‌گیری F را برای روش‌های Ego-Path، کاتز، پیوند دوست و متا-مسیر نشان می‌دهد. اندازه‌گیری F میانگین هارمونیک دو معیار درستی و یادآور است. شکل ۵ (الف) و (ب) به ترتیب نتایج این معیار را برای دو مجموعه داده توئیتر و فیسبوک نشان می‌دهد. در این نمودار از مدل طبقه‌بندی DT استفاده شده است. شکل ۵ (ج) و (د)، این نتایج را برای طبقه‌بندی NN و در نهایت شکل ۵ (ه) و (و) این نتایج را بر مبنای طبقه‌بندی NB نشان می‌دهد.

نتایج نشان دهنده برتری روش پیشنهادی Ego-Path نسبت به سایر روش‌ها در مدل طبقه‌بندی DT می‌باشد. نتایج روش پیشنهادی در طبقه‌بندی NN بعد از روش متا-مسیر در رتبه دوم قرار دارد. اما دو روش متا-مسیر و Ego-Path که در ساختار خود از مدل طبقه‌بندی NB استفاده می‌کنند نسبت به دو روش کاتز و پیوند دوست عملکرد ضعیفی دارند. به نظر می‌رسد دو روش دو روش متا-مسیر و Ego-Path که بر پایه مدل طبقه‌بندی و استخراج ویژگی می‌باشد، به شدت به سیاست انتخاب ویژگی‌ها و نوع روش طبقه‌بندی در پیشنهاد لینک حساس هستند.

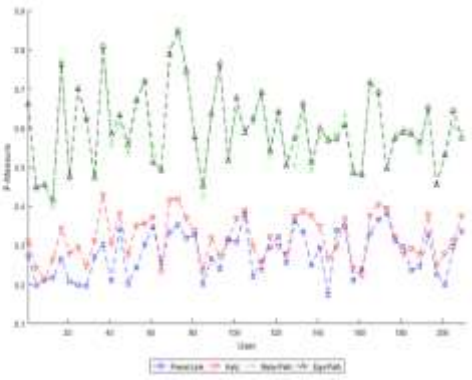
در شکل ۶ نتایج مقایسه تعداد کاربران پیشنهادی مرتبط با تعداد کاربران دوست واقعی برای روش پیشنهادی و سایر روش‌ها گزارش شده است. در اینجا برای هر کاربر هدف به صورت جداگانه، تعداد دوستانی که به درستی تشخیص داده شده‌اند نسبت به تعداد دوستان واقعی در نظر می‌گیریم. شکل ۶ (الف) و (ب) به ترتیب نتایج مقایسه را برای دو مجموعه داده توئیتر و فیسبوک وقتی از مدل طبقه‌بندی DT استفاده می‌کنیم، نشان می‌دهد. شکل ۶ (ج) و (د)، این نتایج را برای طبقه‌بندی NN و در نهایت شکل ۶ (ه) و (و)، نتایج مقایسه را بر مبنای طبقه‌بندی NB نشان می‌دهد. عملکرد روش پیشنهادی از روش‌های دیگر وقتی مدل طبقه‌بندی DT استفاده می‌شود، بهتر است.

نتایج حاصل از آزمایش‌ها دقت بالای روش پیشنهادی را نسبت به روش‌های مورد مقایسه در طبقه‌بند DT نشان می‌دهد. عملکرد بهتر روش پیشنهادی به دلیل استفاده از ویژگی‌های موثر استخراج شده و درنظر گرفتن ساختار توپولوژیک گراف می‌باشد. همچنین با توجه به این موضوع که مردم تمایل به ایجاد روابط جدید با کسانی را دارند که بر روی یک گراف اجتماعی از نظر خصوصیات به آنها

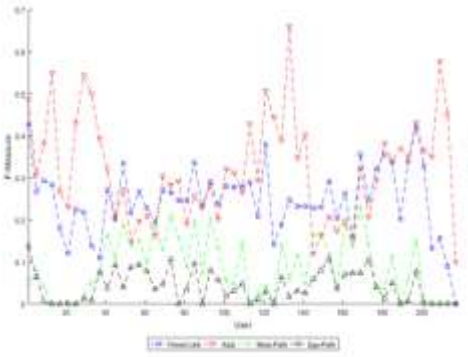
نزدیک‌تر هستند، استفاده از گروه‌ها و روابط بین گروه‌ها در دو ویژگی IEP و OEP به افزایش دقت روش پیشنهادی کمک کرده است. همچنین در نظر گرفتن توئیت‌ها و پیام‌های مشترک بین کاربران در افزایش کارایی روش پیشنهادی نقش موثری داشته است.



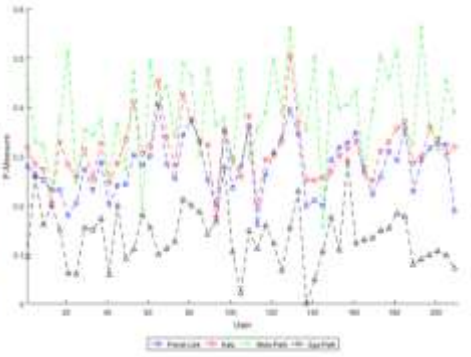
ب: مجموعه داده فیسبوک، مدل طبقه‌بندی DT



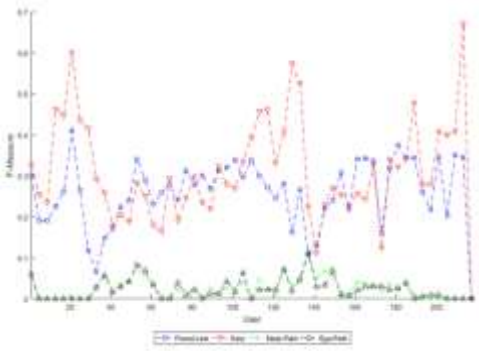
الف: مجموعه داده توئیتر، مدل طبقه‌بندی DT



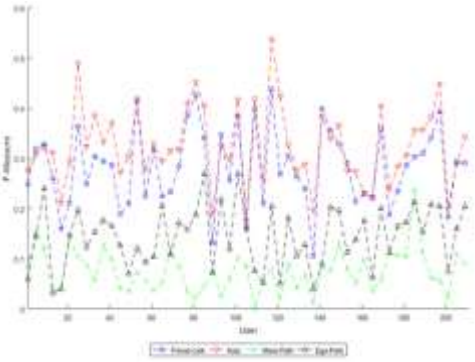
د: مجموعه داده فیسبوک، مدل طبقه‌بندی NN



ج: مجموعه داده توئیتر، مدل طبقه‌بندی NN

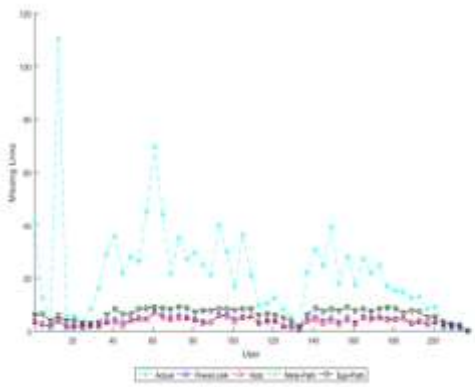


و: مجموعه داده فیسبوک، مدل طبقه‌بندی NB

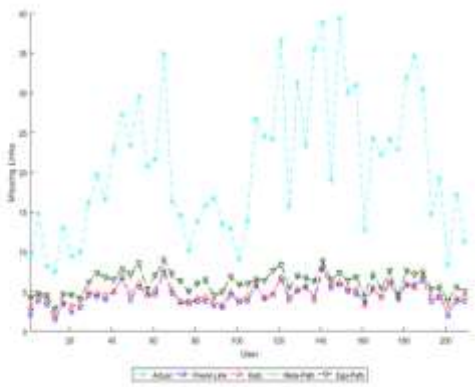


ه: مجموعه داده توئیتر، مدل طبقه‌بندی NB

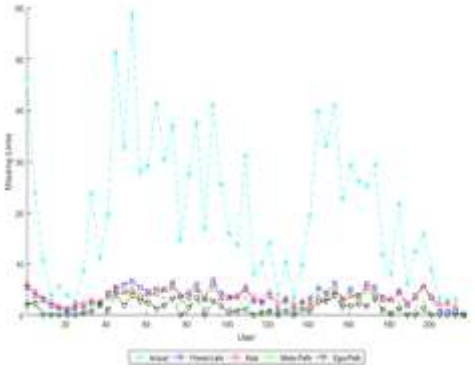
شکل ۵ : نتایج معیار اندازه‌گیری F در روش‌های Ego-Path، کاتز، پیوند دوست و متا-مسیر



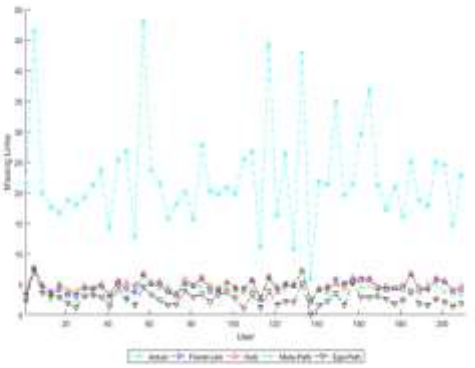
ب: مجموعه داده فیسبوک، مدل طبقه‌بندی DT



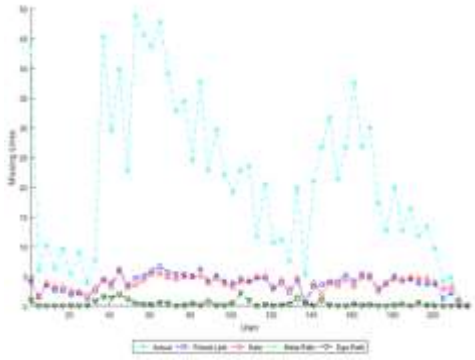
الف: مجموعه داده توئیتر، مدل طبقه‌بندی DT



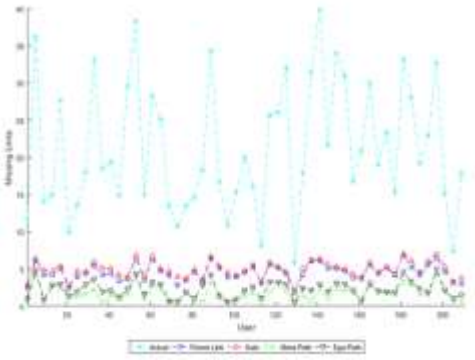
د: مجموعه داده فیسبوک، مدل طبقه‌بندی NN



ج: مجموعه داده توئیتر، مدل طبقه‌بندی NN



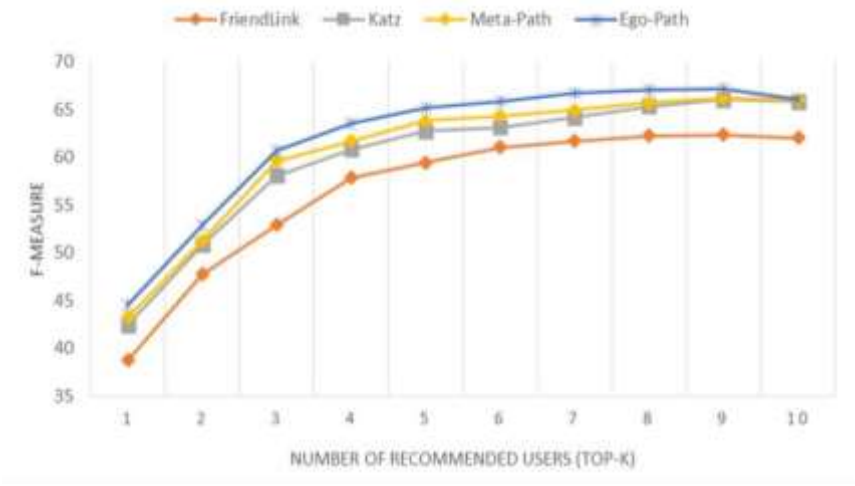
و: مجموعه داده فیسبوک، مدل طبقه‌بندی NB



ه: مجموعه داده توئیتر، مدل طبقه‌بندی NB

شکل ۶ : مقایسه تعداد کاربران صحیح پیشنهادی با داده‌های واقعی

به طور کلی، هر چه تعداد کاربران پیشنهادی $top-k$ بالا باشد، احتمال ظاهر شدن کاربران مرتبط واقعی در زیرمجموعه کاربران پیشنهادی افزایش می‌یابد. این امر باعث کاهش معیار درستی و افزایش معیار یادآور خواهد شد. با این شرایط نیاز به معیارهای ترکیبی نظیر اندازه‌گیری F به منظور بررسی دقیق‌تر مدل پیشنهادی ضروری می‌باشد. شکل ۷ نتایج مقایسه معیار اندازه‌گیری F را برای روش‌های مختلف بر مبنای تعداد پیشنهادات مختلف نشان می‌دهد.



شکل ۷ : عملکرد روش‌های مختلف در معیار اندازه‌گیری F بر مبنای تعداد پیشنهادات مختلف

معیار اندازه‌گیری F برای ارزیابی روش‌ها از میانگین هارمونیک دو معیار درستی و یاداور استفاده می‌کند که معیار مناسبی برای ارزیابی مدل‌های پیش‌بینی لینک است. نتایج این معیار نشان می‌دهد که روش پیشنهادی نسبت به سایر روش‌ها از عملکرد بهتری برخوردار است. در بهترین حالت روش پیشنهادی به دقت % 69 در معیار اندازه‌گیری F با تعداد ۹ کاربر پیشنهادی رسیده است. در رتبه‌های بعدی به ترتیب روش‌های Meta-Path، Katz و FriendLink قرار دارند.

نتایج عددی معیارهای درستی، یادآور، F و میزان برتری در جدول ۲ برای دو مجموعه داده توئیتر و فیسبوک نشان داده شده است. معیار برتری بیانگر درصد موفقیت هر روش نسبت به سایر روش‌ها برای تمام کاربران E^P می‌باشد. نتایج روش پیشنهادی در مقایسه با روش‌های متا-مسیر، کاتز و پیوند دوست با توجه به معیارهای مختلف برای دو مجموعه داده توئیتر و فیسبوک در طبقه‌بند DT عملکرد بهتری را نشان می‌دهد.

جدول ۲ : نتایج ارزیابی روش‌های مختلف با معیارهای درستی، یادآور، اندازه‌گیری F و برتری روی مجموعه داده‌های توئیتر و فیسبوک

روش‌ها	طبقه‌بند	مجموعه داده توئیتر				مجموعه داده فیسبوک			
		برتری	درستی	یادآور	F	برتری	درستی	یادآور	F
کاتز	DT	۰/۰۲۲	۰/۵۵۱	۰/۳۹۵	۰/۴۶۰	۰/۰۰۰	۰/۴۷۰	۰/۳۸۸	۰/۴۲۵
	NN	۰/۱۰۹	۰/۵۴۸	۰/۳۹۴	۰/۴۵۹	۰/۲۸۹	۰/۴۶۷	۰/۳۸۲	۰/۴۲۰
	NB	۰/۳۰۸	۰/۵۵۴	۰/۳۹۹	۰/۴۶۴	۰/۳۱۲	۰/۴۷۵	۰/۳۸۹	۰/۴۲۸
پیوند دوست	DT	۰/۰۷۹	۰/۵۱۴	۰/۲۹۹	۰/۳۷۸	۰/۰۰۱	۰/۴۴۴	۰/۲۴۹	۰/۳۱۹
	NN	۰/۱۸۳	۰/۵۱۳	۰/۲۹۹	۰/۳۷۸	۰/۴۰۶	۰/۴۴۶	۰/۲۴۷	۰/۳۱۸
	NB	۰/۳۸۱	۰/۵۱۷	۰/۲۹۹	۰/۳۷۹	۰/۵۷۹	۰/۴۴۶	۰/۲۴۸	۰/۳۱۹
متا-مسیر	DT	۰/۰۹۸	۰/۹۱۶	۰/۵۶۱	۰/۶۹۶	۰/۰۱۷	۰/۹۸۳	۰/۶۳۰	۰/۷۶۸
	NN	۰/۴۰۴	۰/۷۴۹	۰/۳۳۹	۰/۴۶۷	۰/۱۶۴	۰/۳۰۲	۰/۰۶۲	۰/۱۰۳
	NB	۰/۰۳۷	۰/۲۵۰	۰/۰۵۰	۰/۰۸۴	۰/۰۲۵	۰/۰۷۴	۰/۰۲۱	۰/۰۳۳
Ego-Path	DT	۰/۸۰۰	۰/۹۱۹	۰/۵۶۲	۰/۶۹۸	۰/۹۸۱	۰/۹۸۴	۰/۶۳۰	۰/۷۶۸
	NN	۰/۳۰۱	۰/۴۰۳	۰/۰۹۹	۰/۶۵۹	۰/۱۳۹	۰/۱۶۰	۰/۰۲۳	۰/۰۴۱
	NB	۰/۲۷۲	۰/۴۱۰	۰/۱۰۳	۰/۶۶۵	۰/۰۸۲	۰/۰۶۶	۰/۰۲۱	۰/۰۳۲

با توجه به نتایج حاصل شده در 80 درصد مواردی که از مجموعه داده توئیتر و مدل طبقه‌بندی DT استفاده شده است، روش Ego-Path نسبت به سایر روش‌ها در محاسبه معیار درستی عملکرد بهتری داشته است. این برتری برای مجموعه داده فیسبوک 98.1 درصد است. آزمایش‌ها برتری 0.3 درصدی Ego-Path نسبت به روش متا-مسیر، برتری 40.5 درصدی نسبت پیوند دوست و برتری 36.8 درصدی نسبت به روش کاتز در مجموعه داده توئیتر را نشان می‌دهد. همچنین نتایج برتری 0.1 درصدی Ego-Path نسبت به روش متا-مسیر، برتری 54 درصدی نسبت پیوند دوست و برتری 51.4 درصدی نسبت به روش کاتز در مجموعه داده فیسبوک را نشان می‌دهد. این مقایسه نسبت به معیار درستی و مدل طبقه‌بندی DT محاسبه شده است.

به منظور بررسی دو مدل مبتنی بر طبقه‌بندی (Ego-Path و متا-مسیر)، آزمایش دیگری انجام شده است. در این آزمایش سه معیار دقت، MAE و RMSE مورد در سه طبقه‌بند DT، NN و NB بررسی می‌شود. نتایج این بررسی برای دو مجموعه داده توئیتر و فیسبوک در جدول ۳ نشان داده شده است. این نتایج بر مبنای خروجی مدل‌های طبقه‌بندی برای دو روش Ego-Path و متا-مسیر ارزیابی شده است.

جدول ۳ : نتایج ارزیابی مدل‌های طبقه‌بندی برای روش‌های Ego-Path و متا-مسیر در مجموعه داده توئیتر

روش‌ها	مدل طبقه‌بندی	مجموعه داده توئیتر			مجموعه داده فیسبوک		
		دقت	MAE	RMSE	دقت	MAE	RMSE
متا-مسیر	DT	۰/۹۹۷۸	۰/۰۰۲۱	۰/۰۴۵۹	۰/۹۹۸۷	۰/۰۰۱۲	۰/۰۳۵۲
	NN	۰/۹۹۲۹	۰/۰۰۷۰	۰/۰۸۳۷	۰/۹۹۳۰	۰/۰۰۶۹	۰/۰۸۳۳
	NB	۰/۹۸۵۶	۰/۰۱۴۳	۰/۱۱۸۶	۰/۹۸۷۳	۰/۰۱۲۶	۰/۱۱۱۹
Ego-Path	DT	۰/۹۹۷۹	۰/۰۰۲۰	۰/۰۴۴۷	۰/۹۹۸۹	۰/۰۰۱۲	۰/۰۳۵۲
	NN	۰/۹۹۱۹	۰/۰۰۸۰	۰/۰۸۹۶	۰/۹۹۲۰	۰/۰۰۷۹	۰/۰۸۹۱
	NB	۰/۹۸۸۰	۰/۰۱۱۹	۰/۱۰۸۷	۰/۹۸۶۷	۰/۰۱۳۲	۰/۱۱۴۷

نتایج دقت و میزان خطا مناسبی را برای هر دو روش Ego-Path و متا-مسیر در هر سه طبقه‌بند نشان می‌دهد. در آزمایش‌های قبلی دو طبقه‌بند NN و NB عملکرد ضعیفی در روش پیشنهادی داشتند، اما در اینجا میزان دقت ارائه شده به مراتب عملکرد بالای روش پیشنهادی را نشان می‌دهد. دلیل این است که نتایج این آزمایش بر مبنای روش‌های طبقه‌بندی و وابسته به نمونه‌های مجموعه داده است. در واقع تعداد نمونه‌هایی با کلاس منفی (بدون ارتباط) در مجموعه داده‌های ارائه شده برای شبکه‌های اجتماعی بسیار بیشتر از نمونه‌هایی با کلاس مثبت است. بررسی‌ها نشان می‌دهد که تعداد نمونه‌های منفی 80 تا 90 درصد کل نمونه‌ها را شامل می‌شود. این امر به دلیل عدم وجود ارتباط بین یک کاربر و بسیاری از کاربران دیگر است. بنابراین نتایج حاصل شده از معیارهای در این آزمایش بیشتر مبتنی بر تشخیص نمونه‌های منفی است. اما به طور کلی روش پیشنهادی عملکرد بهتری نسبت به روش متا-مسیر در طبقه‌بند DT نشان می‌دهد، هر چند این برتری محسوس نیست.

۵- نتیجه‌گیری و پیشنهادات آتی

در این تحقیق بررسی و مقایسه جدیدترین روش‌های ارائه شده در چند سال اخیر جهت پیشنهاد لینک در شبکه‌های اجتماعی انجام شده است. با تجربیات حاصل شده از بررسی‌های انجام شده یک مدل جدید پیش‌بینی لینک با عنوان Ego-Path که مبتنی برای شبکه‌های خود محور می‌باشد ارائه شده است. به منظور ارائه مدلی جهت پیش‌بینی لینک از خصوصیات روش‌های طبقه‌بندی استفاده شده است که در آن مسئله پیش‌بینی لینک به یک مسئله طبقه‌بندی با دو کلاس مثبت و منفی تبدیل می‌کند، جائیکه کلاس مثبت نشان دهنده ارتباط و کلاس منفی نشان دهنده عدم ارتباط دو کاربر است. همچنین برای محاسبه شباهت بین دو کاربر u و v از ویژگی‌های اعتبار، خوش‌بینی، تعداد همسایه‌های مشترک، تعداد مسیر با طول‌های متفاوت، تعداد توئییت‌های مشترک، تعداد مسیرهای خود محور داخلی و تعداد مسیرهای خود محور خارجی استفاده شد. برای کارهای آتی پیشنهاد می‌شود معیارهای شباهت بین دو کاربر را با توجه به پروفایل کاربر تنظیم کرد و از ویژگی‌هایی نظیر سن کاربر، موقعیت مکانی، شغل و غیره در محاسبه میزان شباهت بهره گرفت.

مراجع

[1] D. Liben-Nowell and J.Kleinberg, "The link-prediction problem for social networks," *journal of the Association for Information Science and Technology*, vol.58, no. 7, pp.1019-1031, 2007.

[2] T. E. Webster, "The Need to Engage With Smartphones and Social Network Sites (SNSs) at Korean Universities," *Recent Developments in Technology-Enhanced and Computer-Assisted Language Learning*, pp. 122–143, 2020.

[3] K. Li, L. Tu, and L. Chai, "Ensemble-model-based link prediction of complex networks," *Computer Networks*, vol. 166, pp. 106978, 2020.

[4] Y. Sada, J. Hou, P. Richardson, H. El-Serag, and J. Davila, "Validation of Case Finding Algorithms for Hepatocellular Cancer From Administrative Data and Electronic Health Records Using Natural Language Processing," *Medical Care*, vol. 54, no. 2, 2016.

[5] S.Samanta, and M. Pal, "Link Prediction in Social Networks," *Graph Theoretic Approaches for Analyzing Large-Scale Social Networks*, 164. 2017

[6] S.Rafiee, C.Salavati, and A. Abdollahpouri, "CNDP: Link prediction based on common neighbors degree penalization," *Physica A: Statistical Mechanics and its Applications*, vol.539, pp.122950, 2020.

[7] D. Liben-Nowell, and J. Kleinberg, "The Link-prediction Problem for Social Networks," *Journal of the American Society for Information Science and Technology*, vol.58, no.7, pp.1019-1031, 2007.

[8] H. Mohammad, "Link Prediction In Social Networks, Indiana university," *Social Network Data Analytics*, pp. 243-275, 2011.

[9] M.Jalili, Y.Orouskhani, M.Asgari, N.Alipourfard, and M. Perc, "Link prediction in multiplex online social networks," *Royal Society Open Science*, vol.4, no.2, pp.160863, 2017.

[10] W.Cui, C.Pu, Z.Xu, S.Cai, J.Yang, and A.Michaelson, , "Bounded link prediction in very large networks," *Physica A: Statistical Mechanics and its Applications*, vol.457, pp.202-214, 2016.

[11] F.Parvazeh, A.Harounabadi, and M. A.Naizari, "A Recommender System for Making Friendship in Social Networks Using Graph Theory and users profile," *Journal of Current Research in Science*, no.1, pp.535, 2016.

[12] F. Liu, B.Liu, C.Sun, M.Liu, and X. Wang, "Deep belief network-based approaches for link prediction in signed social networks," *Entropy*, vol.17, no.4, pp.2140-2169, 2015.

[13] R.Laishram, K.Mehrotra, and C. K. Mohan, "Link Prediction in Social Networks with Edge Aging," *In proc.In Tools with Artificial Intelligence (ICTAI),* November 2016, pp. 606-613.

[14] M.Naderipour,S.Bastani, and M. F.Zarandi, "A Type-2 Fuzzy Model for Link Prediction in Social Network. World Academy of Science, Engineering and Technology," *International Journal of Computer, Electrical, Automation, Control and Information Engineering*," vol.10, no.7, pp.1355-1360, 2016.

[15] P. L. Szczepański, A. S.Barcz, T.P.Michalak, and T. Rahwan, "The game-theoretic interaction index on social networks with applications to link prediction and community detection," *In proc Twenty-Fourth International Joint Conference on Artificial Intelligence* , June 2015, pp. 638-644.

[16] J.Zhao, L. Miao, J.Yang, H.Fang, Q. M.Zhang, M.Nie, and T. Zhou, "Prediction of links and weights in networks by reliable routes," *Scientific reports*, vol.5, pp.12261, 2015.

[17] S.Han, and Y. Xu, "Link Prediction in Microblog Network Using Supervised Learning with Multiple Features, " *in JCP*, vol.11, no.1, pp.72-82, 2016.

[18] M.Zhang, , Z.Cui, , S.Jiang, and Y.Chen, "Beyond Link Prediction: Predicting Hyperlinks in Adjacency Space, " *AAAI-2018* , pp. 4430-4437, 2018.

[19] Z.Huo, X.Huang, and X. Hu, "Link Prediction with Personalized Social Influence," *in proc.Conference on Artificial* Intelligence, 2018, pp. 136-139.

[20] P.Pei, B.Liu, and L. Jiao, "Link prediction in complex networks based on an information allocation index," *Physica A: Statistical Mechanics and its Applications*, vol.470, pp.1-11, 2017.

[21] J.Leskovec, and J.Mcauley, "Learning to discover social circles in ego networks," *In Advances in neural information processing systems*, pp. 539-547, 2012.

[22] K. K.Shang, , M.Small, , X. K.Xu, and W. S. Yan, "The role of direct links for link prediction in evolving networks," *EPL (Europhysics Letters),* vol.117, no.2, pp.28002, 2017.

[23] S.Sharma, and A.Singh, "An efficient method for link prediction in weighted multiplex networks," *Computational Social Networks*, vol.3, no. 1, pp.7, 2016.

[24] A.Anagnostopoulos, J.Łącki, S. Lattanzi, S. Leonardi, and M. Mahdian, "Community detection on evolving graphs," *In Advances in Neural Information Processing Systems*, pp. 3522-3530, 2016.

[25] A.Papadimitriou, P.Symeonidis, and Y.Manolopoulos, "Fast and accurate link prediction in social networking systems," *Journal of Systems and Software*, vol.85, no. 9, pp. 2119-2132, 2012

[26] L.Bai, X.Cheng, J. Liang, and Y.Guo, "Fast graph clustering with a new description model for community detection," *Information Sciences*, vol. 388- 389, pp. 37-47, 2017.

[27] F.Aghabozorgi, and M.R.Khayyambashi, "A new similarity measure for link prediction based on local structures in social networks," *Physica A: Statistical Mechanics and its Applications*, vol.501, pp.12-23. 2018

[28] X.Cao, Y.Zheng, C.Shi, J.Li, & B.Wu, "Meta-path-based link prediction in schema-rich heterogeneous information network," *International Journal of Data Science and Analytics*, vol.3, no.4, pp.285-296, 2017.

Use of Classification Models for Optimize Link Prediction in the Ego-Social Networks

Soheila Nemati [1], MehdiSadeghzadeh[2*], maziyar ganjoo[3]

[1]Department of Computer Engineering, Faculty of Engineering, Bushehr Branch, Islamic Azad University, Bushehr, Iran
[2*]Department of Computer Engineering, Faculty of Engineering, Mahshahr Branch, Islamic Azad University, Mahshahr, Iran
[3]Department of Information Technology Engineering, Faculty of Engineering, Bushehr Branch, Islamic Azad University, Bushehr, Iran
1: s.nemati2012@yahoo.com
2*: sadegh_1999@yahoo.com
3: ganjoo@gmail.com

ABSTRACT:

Social propositional systems are a new generation of systems that use the social network as a user modeling platform to maximize some challenges by using rich interactive data volumes. To make Social networking sites offer new friends to registered users based on local graph features. The main purpose of the link prediction problem on social networks is to suggest a list of users to a particular user that they will probably be communicating in the future. In this research, a prediction method for the link is presented based on the characteristics of classification models. Here, the prediction problem of the link is transformed into a classifying problem with two positive and negative classes, where the positive class represents the relationship and the negative class indicates that the two users are not communicating. Three classical classes DT, NN and NB are used for classification work. To create the dataset, the features of credibility, optimism, number of neighbors, the number of paths of different lengths, the number of shared tweets, the number of internal and external axes are used. Although self-centered grids do not have much overlap in the rings, experiments show that the consideration of self-directed pathways significantly improves predictive performance. The DT classification has recorded the best performance with an average accuracy of 99.85%.

KEYWORDS: Ego-social networks, Similarity criterion, Data classification, Feature extraction