مقاله پژوهشی

افزایش دقت شناسایی جوامع در شبکه‌های اجتماعی با بهبود رویکرد انتشار برچسب

نفیسه افخمی[۱] | نازبانو فرزانه[۲] | حسن شاکری*[۳]

[۱] گروه مهندسی کامپیوتر، دانشگاه بین‌المللی امام رضا، مشهد، ایران، nafiseafkhami64@gmail.com

[۲] گروه مهندسی کامپیوتر، دانشگاه بین‌المللی امام رضا، مشهد، ایران، nazbanou@farzaneh@gmail.com

[۳] گروه مهندسی کامپیوتر، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران ، hassanshakeri@iau.ac.ir

نویسنده مسئول

*حسن شاکری، دانشیار، گروه مهندسی کامپیوتر، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران، hassanshakeri@iau.ac.ir

تاریخ دریافت:
تاریخ بازنگری: ۲۱ فروردین ۱۴۰۴
تاریخ پذیرش:

چکیده:

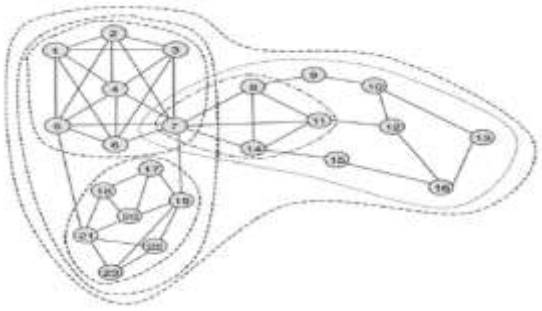
شناسایی جوامع در شبکه‌های بزرگ یک موضوع پرکاربرد در تحلیل شبکه‌های اجتماعی است و ارائه الگوریتم با دقت و کارایی مطلوب برای استخراج جوامع اهمیت زیادی دارد. رویکردهای مختلفی برای شناسایی جامعه‌ها وجود دارد. ازجمله می‌توان به رویکرد انتشار برچسب اشاره کرد که در آن ابتدا مهم‌ترین رأس‌های شبکه برمبنای معیارهای مرکزیت تعیین می‌شوند و برچسب‌های جامعه متفاوت به آنها انتساب داده می‌شود. سپس برچسب هر یک از این رأس‌ها به رأس‌های اطراف آنها انتشار می‌یابد. هدف این پژوهش، بهبود یک الگوریتم شناسایی جامعه موسوم به LBLD است. این الگوریتم ابتدا براساس یک معیار شباهت، تعدادی از مهم‌ترین رأس‌های شبکه را تعیین می‌کند. سپس با یک رویکرد متوازن، جوامع توسعه داده می‌شوند و در نهایت یک فاز ادغام اجرا می‌شود تا جوامع کوچک با یکدیگر ترکیب شوند. ایده پیشنهادی ما استفاده از یک معیار الهام‌گرفته از مفهوم h-index برای بهبود دقت تشخیص جوامع است به این ترتیب که زیرگراف‌هایی به عنوان جامعه شناسایی شوند که حداقل pدرصد از رأس‌های آنها درجه حداقل k داشته باشد. اعمال روش پیشنهادی بر روی مجموعه‌داده‌های شناخته‌شده در این حوزه و و مقایسه نتایج نشان می‌دهد که روش پیشنهادی نسبت به روش‌های مشابه باعث بهبود دقت در استخراج جوامع شده است.

کلید واژه‌ها: شناسایی جوامع، مدل‌سازی شبکه‌های اجتماعی، الگوریتم انتشار برچسب، خوشه‌بندی.

۱-مقدمه

رشد سریع رسانه‌ها و شبکه‌های اجتماعی تأثیر شگرفی بر ابعاد مختلف زندگی انسان‌ها گذاشته است به طوری که مدل‌سازی و تحلیل شبکه‌های اجتماعی به یک مبحث مطرح در تحقیقات و نیز در کاربردهای مختلف تبدیل شده است. مدل‌سازی و تحلیل شبکه‌های اجتماعی زمینه‌های مختلفی را شامل می‌شود که از جمله می‌توان به تعیین مرکزیت و اهمیت رأس‌های گراف شبکه، پیش‌بینی رفتار کاربران، ، طبقه‌بندی کاربران و شناسایی جوامع اشاره کرد و در حوزه‌های مختلفی مانند بهبود کارآمدی سیستم‌های پیشنهاددهنده ، تحلیل‌های زیست‌شناسی از قبیل بررسی شبکه پروتئین-پروتئین و مدل‌سازی شبکه‌های ارجاع کاربرد دارد [۱]. برای مدل‌سازی شبکه‌های اجتماعی، از مفاهیم نظریه گراف استفاده می‌شود. برای این منظور شبکه اجتماعی به صورت یک گراف بازنمایی می‌شود که اعضای شبکه به عنوان رأس‌های گراف و روابط بین آنها (ازقبیل روابط دوستی یا دنبال‌کردن) به عنوان یال‌ها یا لبه‌های گراف در نظر گرفته می‌شود.

شناسایی جوامع یکی از مهم‌ترین و پرکاربردترین حوزه‌های مدل‌سازی و تحلیل شبکه‌های اجتماعی است. هر جامعه عبارت است از گروهی از اعضای شبکه‌های اجتماع (یا به عبارت دیگر یک زیرگراف از گراف اصلی) که عناصر آن شباهت زیادی از نظر ویژگی‌ها و ارتباطات زیادی با یکدیگر دارند (شکل ۱). در واقع، مبحث شناسایی جوامع معادل مفهوم خوشه‌بندی است که به صورت کلی در مباحث داده‌کاوی مطرح است اما در شبکه‌های اجتماعی، به جای اصطلاح خوشه‌بندی، عمدتاً از اصطلاح «شناسایی جوامع» استفاده می‌شود و ضمناً در شناسایی جوامع، علاوه بر رویکردها و تکنیک‌های خوشه‌بندی کلاسیک از مفاهیم شبکه‌های اجتماعی به ویژه ارتباطات بین کاربران شبکه به‌صورت گسترده استفاده می‌شود. یک معیار مهم در کیفیت تفکیک شبکه به جوامع مختلف این است که اتصالات و لینک‌ها در داخل هر جامعه تا حدامکان زیاد ولی در بین هر دو جامعه مختلف، تا حد امکان کم باشد. استخراج جوامع به درک ساختار آنها و ارتباطات بین گره‌ها و نحوه کارکردن شبکه کمک می‌کند [۲].



شکل ۱. استخراج جوامع در شبکه‌های اجتماعی [۲]

رویکردهای مختلفی برای شناسایی جامعه‌ها در شبکه‌های اجتماعی وجود دارد که ازجمله می‌توان به روش‌های مبتنی بر الگوریتم‌های کلاسیک خوشه‌بندی براساس معیارهای شباهت در ویژگی‌ها اشاره کرد. ازجمله الگوریتم K-means به عنوان یکی از معروف‌ترین الگوریتم‌های کلاسیک خوشه‌بندی با معیارهای شباهت مختلف از قبیل شباهت کسینوسی و یا شباهت جاکارد در استخراج جوامع در شبکه‌های اجتماعی نیز مورد استفاده قرار می‌گیرند. دسته‌ای دیگر از روشها بر یافتن زیرگراف‌های با ارتباطات داخلی زیاد مبتنی هستند که از جمله می‌توان به کشف جوامع مبتنی بر مفهوم Clique اشاره کرد. یک رویکرد جدید برای استخراج جوامع، انتشار برچسب است که یک تکنیک سریع برای توسعه جوامع اولیه به گره‌های اطراف آنها است. به عنوان نمونه می‌توان به الگوریتم ارائه‌شده در مرجع [۳] به نام LSMD[1] اشاره کرد که یک الگوریتم انتشار چندسطحی شباهت محلی است و براساس این رویکرد عمل می‌کند.

با وجود ارائه کارهای تحقیقاتی مختلف در حوزه استخراج جوامع در شبکه‌های اجتماعی، همچنان افزایش دقت و کاهش مرتبه زمانی الگوریتم‌ها در این زمینه به عنوان مسائل باز تحقیقاتی مطرح هستند [۱]. به همین دلیل در این تحقیق ما قصد داریم راهکاری برای بهبود یک الگوریتم شناسایی جوامع معرفی‌شده در [۱] موسوم به الگوریتم LBLD[2] ارائه کنیم. الگوریتم LBLD ابتدا براساس یک معیار شباهت، پنج‌درصد از مهم‌ترین رأس‌های شبکه را تعیین می‌کند. سپس با یک رویکرد متوازن، جوامع هم از مرکز و هم از مرزها توسعه داده می‌شوند و در نهایت یک فاز ادغام اجرا می‌شود تا جوامع کوچک با یکدیگر ترکیب شوند و جوامع مطلوب را تشکیل دهند.

در این تحقیق یک راهکار بهبودیافته براساس روش ارائه‌شده در مرجع [۱] پیشنهاد می‌کنیم. ایده پیشنهادی ما استفاده از یک معیار الهام‌گرفته از مفهوم h-index برای بهبود دقت تشخیص جوامع است. به این ترتیب که زیرگراف‌هایی به عنوان جامعه شناسایی شوند که حداقل تعداد n رأس آنها درجه حداقل k داشته باشد. رویکرد کلی مورد استفاده در روش پیشنهادی، مشابه روش پایه، انتشار برچسب جامعه از گره‌های مهم به گره‌های مجاور آنها است. با وجود این تغییرات و نوآوری‌های عمده‌ای در روش پیشنهادی نسبت به الگوریتم پایه اعمال شده که باعث افزایش دقت شناسایی جوامع شده است. مهم‌ترین نوآوری‌ها و تغییرات پیشنهادی نسبت به روش پایه عبارتند از:

[1] Local Similarity Multi-level Diffusion
[2] Local Balanced Label Diffusion

۱. استفاده از معیار جاکارد برای تعیین میزان شباهت گره‌ها از نظر همسایه‌های مشترک و استفاده از آن در تعیین میزان اهمیت گره‌ها که براساس آن، جوامع اولیه شکل می‌گیرند. دلیل استفاده از معیار شباهت جاکارد، کارآمدی آن در مسائل مبتنی بر گراف و مشخصاً حوزه مدل‌سازی و تحلیل شبکه‌های اجتماعی است که تأثیر آن در ارزیابی روش پیشنهادی در بخش ۴ نیز مشهود است.

۲. استفاده از یک معیار الهام‌گرفته از معیار h-index مربوط به رتبه‌بندی محققان برای انتخاب نهایی جوامع.

در ادامه این مقاله در بخش ۲ کارهای تحقیقاتی مرتبط در این حوزه مورد بررسی قرار می‌گیرد.. در بخش ۳ به توصیف روش پیشنهادی جهت افزایش دقت شناسایی جوامع در شبکه‌های اجتماعی پرداخته می‌شود. بخش ۴، ارزیابی روش پیشنهادی و مقایسه آن با کارهای پژوهشی پیشین را ارائه می‌دهد و در نهایت، در بخش ۵ نتیجه‌گیری و ایده‌های پیشنهادی برای کارهای آینده مطرح می‌گردد.

۲- پیشینه تحقیق

با توجه به اهمیت شناسایی جوامع به عنوان یک ابزار ضروری برای تحلیل اطلاعات پنهان شبکه‌ها، الگوریتم‌های متنوعی برای این منظور ارائه شده است که ازجمله می‌توان به رویکردهای محلی، مبتنی بر ماژولاریتی و مبتنی بر تجزیه ماتریس اشاره کرد. یکی از روش‌های ارائه‌شده برای کشف جوامع، LP-LPA[1] نام دارد که مقدار قدرت لینک‌ها را برای یال‌ها و مقدار تأثیرگذاری برچسب‌ها را برای رأس‌استفاده می‌کند تا از انتخاب تصادفی حالت‌های قطع اتصال جلوگیری کند. روش ارائه‌شده در [۴] موسوم به LPA[2] یا NI-LPA[3] از ترکیب اندیس جاکارد و انتشار سیگنال بین گره‌ها استفاده می‌کند تا امتیاز اهمیت دقیق‌تری به رأس‌ها انتساب دهد. به طور کلی روش‌های مبتنی بر LPA اغلب بر تعریف ملاک‌های اهمیت، مرتب‌سازی رأس‌ها بر مبنای میزان اهمیت آنها و تعریف معیارهایی برای حالت‌های شکستن اتصال مبتنی هستند. اما این روش‌ها هم با مشکلاتی از قبیل پایین‌بودن دقت، مرتبه زمانی بالا و ناپایداری الگوریتم مواجه هستند.

روش CDME[4] که در [۵] توسط سان و همکاران ارائه شده است، یک تکنیک مبتنی بر اثر متیو است که از رویکرد محلی استفاده می‌کند. هدف این روش، مشخص کردن پارامترهای روال‌های بهینه‌سازی است و مزیت عمده این روش در این راستا کاهش زمان این بهینه‌سازی‌ها است. در این تحقیق، شبکه به عنوان یک سیستم اجتماعی در نظر گرفته شده است و با الهام‌گیری از اثر متیو رویکردی ارائه شده است که جوامع با کیفیت بالا و پویا را استخراج می‌کند. یک مزیت دیگر این الگوریتم این است که به صورت محلی عمل می‌کند و فقط نیاز به محاسبه میزان جذابیت گره‌های همسایه دارد که این ویژگی باعث می‌شود که پردازش شبکه‌های با مقیاس بالا امکان‌پذیر باشد.

LCDR[5] [۶] یک روش محلی دیگر است که براساس رتبه‌بندی گره‌ها با ایده کرشهف یا ذخیره‌سازی بار مبتنی است. در این تحقیق ابتدا مزایا و معایب روش‌های محلی شناسایی جوامع مورد بررسی قرار گرفته است. مهم‌ترین مزیت روش‌های محلی، بالابردن کارایی آنها است. در مقابل، این روش‌ها چند مشکل عمده دارند که پایین‌بودن دقت و مقیاس‌پذیری و ناپایداری جوامع استخراج‌شده جزو مهم‌ترین معایب روش‌های مذکور هستند. در واقع، دقت و کیفیت جوامع در روش‌های محلی وابستگی زیادی به انتخاب گره‌های مهم دارد که به عنوان هسته‌های اولیه جوامع انتخاب می‌شوند. برای حل این مشکلات، در این مرجع یک الگوریتم محلی جدید کشف جوامع معرفی می‌کند که بر رتبه‌بندی گره‌های با اهمیت بالا مبتنی است. در این الگوریتم یک شاخص جدید برای محاسبه اهمیت گره‌ها ارائه شده است. این شاخص می‌تواند با توجه به محلی‌بودن شبکه، اهمیت همه گره‌های شبکه را به صورت کامل بازتاب دهد. الگوریتم LCDR ابتدا گره‌های مهم را برای بسط جوامع اولیه براساس یک معیار شباهت محلی انتخاب می‌کند تا این که همه گره‌ها عضو جوامع شوند. در انتها جوامع شناسایی‌شده ادغام می‌شوند تا جوامع نهایی شکل بگیرند.

---

[1] Label influence Policy for Label Propagation Algorithm
[2] Label Propagation Algorithm
[3] Importance based Label Propagation Algorithm
[4] Community Detection based on the Matthew Effect
[5] Local Community Detection based on high importance nodes Ranking

زارع‌زاده و همکاران در [۷] یک گونه بهبودیافته از الگوریتم موسوم به GCN[1] را ارائه کرده‌اند که از امتیازدهی ترکیبی به گره‌ها و بروزرسانی برچسب‌های همگام برای افزایش دقت تشخیص جوامع استفاده می‌کند. در این مقاله ابتدا مشکلات الگوریتم‌های موجود در راستای افزایش دقت شناسایی جوامع مورد تحلیل قرار گرفته است که مهم‌ترین آنها عبارتند از دقت پایین، تصادفی‌بودن و ناپایداری. برای حل این مشکلات، این تحقیق یک روش بهبودیافته برای شناسایی جوامع پیشنهاد می‌کند که از یک امتیازدهی ترکیبی برای گره‌ها و بروزرسانی همگام برچسب گره‌های مرزی استفاده می‌کند و همچنین مشکل بروزرسانی تصادفی را برطرف می‌کند. برای بروزرسانی برچسب گره‌های مرزی از معیارهای مرکزیت درجه‌ای[2] و همسایه‌های مشترک استفاده می‌شود. همچنین یک روش جدید برای ادغام جوامع پیشنهاد شده است که نسبت به روش‌های مبتنی بر پیمانگی[3] سریع‌تر است.

همچنین یک روش دیگر به نام APAS[4] در [۸] معرفی شده است که از الگوریتم مبتنی بر انتشار استفاده می‌کند که جوامع را با بهره‌گیری از یک ماتریس شباهت نامتقارن تطبیقی شناسایی می‌کند و می‌تواند به صورت کارآمد احتمالات رهبری را بین نقاط داده نشان دهد. ایده اصلی این الگوریتم این است که مقدار شباهت بین دو گره i و k متقارن نیست و بنابراین یک ماتریس شباهت با عنوان ماتریس انتشار وابستگی با شباهت تطبیقی معرفی شده است که می‌تواند به صورت کارا احتمالات رهبری بین نقاط داده را نشان دهد. این ماتریس، شباهت متقارن را تبدیل به شباهت نامتقارن می‌کند که نتیجه آن بهبود کارآمدی الگوریتم شناسایی جوامع در شبکه‌های جهان واقعی بوده است.

همچنین برخی از منابع مانند [۹] و [۱۰] به بررسی کاربردهای شناسایی جوامع در زمینه‌های مختلف پرداخته‌اند. یکی از مهم‌ترین کاربردها که در مرجع [۹] مورد اشاره قرار گرفته است تحلیل یک شبکه از اقلام (کالاها) برای فروش در وب‌سایت‌های فروش آنلاین بزرگ است. اقلام به عنوان یک شبکه درنظر گرفته می‌شوند و در صورتی که توسط یک خریدار واحد خریداری شوند بین آنها لینک برقرار می‌شود. این رویکرد باعث خوشه‌بندی کالاها به صورت مناسب و در نتیجه تأثیر مثبت و مطلوب بر رفتار خرید مشتریان شده است.

در مرجع [۱۰] با بررسی تکنیک‌های مختلف شناسایی جوامع، براساس ویژگی‌های شبکه‌ها راهنمایی‌هایی برای کمک به انتخاب بهترین روش کشف جوامع برای هر شبکه و کاربرد معین ارائه شده است. همچنین قواعد ارائه‌شده در این تحقیق، محدودیت‌های استفاده از هر الگوریتم مشخص برای هر شبکه را با توجه به ویژگی‌های ماکروسکوپی شبکه بیان می‌کند. همچنین از ترکیب پارامترها به عنوان یک شاخص قابل‌اندازه‌گیری برای یافته محدوده‌های اتکاپذیری الگوریتم‌های مختلف استفاده شده است و نیز میزان وابستگی قدرت شناسایی و زمان پردازش هر الگوریتم به اندازه شبکه مورد بررسی قرار گرفته است.

در پژوهش [۱۱]، یک الگوریتم ژنتیک به نام الگوریتم ژنتیک-NET ارائه شد که مفهومی از امتیاز جامعه را برای نشان دادن کیفیت از بخش‌بندی شبکه‌های اجتماعی به کار می‌گیرد. امتیاز جامعه، به حداکثر رساندن ارتباطات داخلی در ساختار جامعه است. این روش جستجوی نامعتبر را زمانی به طور موثر کاهش می‌دهد که فقط همبستگی واقعی تمام گره‌ها در هر عملگر در نظر گرفته شود.

در پژوهش [۱۲]، یک الگوریتم اکتشافی برای بخش‌بندی گراف‌ها ارائه شد. این الگوریتم هنوز هم در ترکیب با تکنیک‌های دیگر به کار می‌رود. هدف آن به حداقل رساندن یک تابع ارزیابی است که تفاوت لینک‌های درون جامعه و بین جامعه است. انگیزه پشت این روش، مسأله تقسیم‌بندی مدارهای الکترونیکی بر روی بوردها بود که در آن رأس‌ها در بوردهای مختلف قرار بود با حداقل تعداد اتصالات به هم متصل شوند. در واقع، تابع بهینه‌سازی برمبنای حداقل تعداد اتصالات بین بوردهای مختلف تعریف می‌شود که این تابع را می‌توان به صورت تفاوت بین تعداد یال‌های داخل مجموعه‌ها و تعداد یال‌های قرار گرفته بین آنها تعریف کرد.

---

[1] Graph Convolutional Network
[2] Degree centrality
[3] modularity-based
[4] Affinity Propagation with Adaptive Similarity

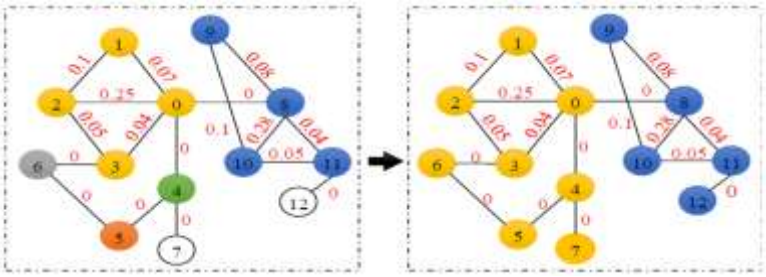
مرجع [۱۳] یک الگوریتم مبتنی بر الگوریتم بهینه‌سازی شاهین هریس و چند تابع تکاملی دیگر به عنوان یک تکنیک یادگیری بهبودیافته ارائه می‌کند که هدف از آنها برقراری تعادل بین استخراج و اکتشاف است. این رویکرد باعث افزایش قابل ملاحظه دقت و صحت شناسایی جوامع در شبکه‌های اجتماعی شده است.

در مرجع [۱۴] یک چارچوب مبتنی بر موقعیت جغرافیایی و اعتماد ارائه شده است که با الگوریتم‌های تشخیص جوامع ترکیب شده است تا کاربران بدخواه را در جوامع شناسایی و فیلتر کند. این عملکرد در شبکه‌های اجتماعی نسل پنجم مورد توجه قرار گرفته است. چارچوب ارائه‌شده از اطلاعات جغرافیایی و اعتماد اجتماعی در داخل شبکه و رویکرد مبتنی بر هوش مصنوعی در کشف جوامع بهره می‌گیرد تا کاربرانی را که باعث آسیب‌رساندن به دیگران می‌شوند شناسایی کند. مزیت این چارچوب نسبت به الگوریتم‌های مشابه این است که خصوصیت‌هایی را که یک کاربر بدخواه و آسیب‌رسان نمی‌توان به راحتی جعل کند، مورد استفاده قرار می‌دهد. مؤلفان این مقاله، چارچوب پیشنهادی خود را بر روی داده‌های شبکه اجتماعی مصنوعی مورد ارزیابی قرار داده‌اند که نتیجه آن بهبود دقت شناسایی کاربرانی بوده است که بالقوه بدخواه هستند.

الگوریتمهای انتشار برچسب جزو روش‌های یادگیری نیمه‌نظارت‌شده و سریع در حوزه شناسایی جوامع هستند. اما ناپایداری و تصادفی‌بودن آنها چالش مهمی محسوب می‌شود. بسیاری از محققان، راهکارهایی برای بهبود تکنیک انتشار برچسب ارائه کرده‌اند که اغلب آنها برای کشف جوامع بزرگ مناسب نیست زیرا پیچیدگی زمانی آنها بسیار بالا است. بنابراین در مرجع [۱۵] یک الگوریتم انتشار برچسب برمبنای معیار h-index تطبیقی ارائه شده است که پایداری و دقت LPA را با استفاده از معیار h-index اصلاح‌شده به عنوان یک ملاک اهمیت گره، بهبود می‌بخشد. با آن که روش پیشنهادی در این پایان‌نامه نیز از ترکیب معیار h-index با رویکرد انتشار برچسب استفاده می‌کند، تفاوت‌های قابل ملاحظه‌ای نسبت به روش مرجع [۱۶] دارد زیرا اولا روش پیشنهادی برمبنای بهبود الگوریتم LBLD طراحی شده و ثانیا در معیار h-index به منظور متناسب‌شدن با کاربرد شناسایی جوامع، اصلاحاتی پیشنهاد شده است که جزئیات این تغییرات و نوآوری‌ها را در فصل بعد که به توصیف روش پیشنهادی می‌پردازد ارائه خواهیم کرد.

کار تحقیقاتی که به عنوان راهکار پایه در این پژوهش مورد مطالعه و بررسی قرار گرفته است در مرجع [۱] ارائه شده است. الگوریتم پیشنهادی در این روش LBLD نام دارد که برگرفته از سرنام کلمات عبارت «انتشار برچسب متوازن محلی» است. این الگوریتم ابتدا براساس یک معیار شباهت، درصدی از گره‌های شبکه را که بیشترین میزان اهمیت را دارند، شناسایی می‌کند. به این گره‌ها برچسب‌های متوالی و متفاوت انتساب داده می‌شود. سپس با یک رویکرد متوازن و با انتشار برچسب از گره‌های اولیه به رأس‌های مجاور آنها، جوامع توسعه داده می‌شوند. در انتها یک مرحله ادغام جوامع اجرا می‌شود تا جوامع کوچک با یکدیگر ترکیب شوند و جوامع با بزرگی مطلوب و در عین حال قوی را تشکیل دهند.

شکل ۲ فرآیند انتشار برچسب در الگوریتم LBLD در یک شبکه نمونه را نشان می‌دهد. در شکل سمت چپ، مهم‌ترین گره‌ها شناسایی شده و برچسب‌های متفاوت به آنها انتساب داده شده که با رنگ‌های متمایز نشان داده شده است. در شکل سمت راست، بروزرسانی برچسب‌ها برمبنای انتشار برچسب انجام شده است.



شکل ۲. نمونه‌ای از فرایند شناسایی جوامع براساس الگوریتم LBLD [۱]

۳- راهکار پیشنهادی

۳-۱- توصیف کلی راهکار پیشنهادی

در این بخش، توصیف کلی ایده پیشنهادی ارائه می‌گردد. سپس در بخش ۳-۲ روش پیشنهادی به تفصیل ارائه و تشریح می‌شود.

در مدل‌سازی و تحلیل شبکه‌های اجتماعی، یکی از مهم‌ترین ملاک‌های تخمین میزان شباهت بین گره‌ها شباهت ساختاری رأس‌ها برمبنای تعداد همسایه‌های مشترک است. برای این منظور، معیارهای شباهت مختلفی مانند شباهت کسینوسی و شباهت جاکارد استفاده می‌شوند. در روش پیشنهادی [۱] معیار دیگری توسط مؤلفان ارائه شده است. اما پیشنهاد ما این است که از معیار جاکارد برای این منظور استفاده شود. دلیل استفاده از معیار شباهت جاکارد این است که این معیار در مسائل حوزه مدل‌سازی و تحلیل شبکه‌های اجتماعی و به طور کلی در ارزیابی شباهت بین گره‌ها در گراف بسیار کارآمد است و نتایج ارزیابی‌های ما در بخش ۴ نیز این موضوع را تأیید می‌کند.

این معیار شباهت بین دو رأس گراف را با تقسیم تعداد همسایه‌های مشترک آن دو بر تعداد همسایه‌های مشترک و غیرمشترک آنها تخمین می‌زند. به عبارت دیگر شباهت جاکارد برای دو رأس A و B در یک شبکه اجتماعی برابر است با تعداد اعضای مجموعه اشتراک همسایه‌های A و B تقسیم بر تعداد اعضای مجموعه اجتماع آنها. رابطه (۱) معیار جاکارد را نشان می‌دهد [۲۰]:

$$\mathrm{Sim_j} = \frac{|\mathrm{N_A} \cap \mathrm{N_B}|}{|\mathrm{N_A} \cup \mathrm{N_B}|} \qquad (۱)$$

از طرف دیگر برای بهبود دقت کشف جوامع، در روش پیشنهادی معیاری برگرفته از معیار h-index استفاده می‌کنیم. h-index یک معیار برای رتبه‌بندی پژوهشگران براساس تعداد ارجاعات به کارهای تحقیقاتی آن است. تعریف این معیار به این صورت است که اگر h-index یک محقق برابر با k باشد به این معنی است که حداقل k تا از مقالات او حداقل k بار مورد ارجاع قرار گرفته‌اند. در واقع هر چه مقالات بیشتری به تعداد بیشتری مورد ارجاع قرار گرفته باشند، اهمیت و تأثیرگذاری پژوهشگر مورد نظر بیشتر است. از این ایده می‌توان برای تعیین میزان قوی‌بودن یک جامعه در شبکه‌های اجتماعی استفاده کرد به این ترتیب که هر چه تعداد بیشتری از رأس‌های (افراد) جامعه تعداد لینک‌ها (همسایه‌های) بیشتری داشته باشند، آن جامعه قوی‌تر است. بنابراین در راهکار پیشنهادی معیاری به نام s_k تعریف می‌شود که اگر برابر با p باشد نشان می‌دهد که حداقل p تا از گره‌های جامعه حداقل k همسایه دارند. ایده پیشنهادی این است که با محاسبه مقدار s_k برای یک جامعه که در مرحله اول شناسایی شده است، جوامعی که s_k آنها از حد معینی کمتر است به جوامع کوچک‌تر و قوی‌تری تفکیک شوند. با دقت در تعریف معیار s_k که در این مقاله معرفی شده است، ملاحظه می‌شود که این مفهوم، عیناً معادل معیار h-index نیست بلکه یک معیار تعمیم‌یافته از h-index است و بنابراین تعریف معیار s_k یکی از نوآوری‌های روش پیشنهادی محسوب می‌شود.

۳-۲ شرح کامل روش پیشنهادی

همان‌طور که در بخش قبلی اشاره شد، راهکار پیشنهادی تا حدودی بر الگوریتم انتشار برچسب با توازن محلی یا LBLD [۱] منطبق است ولی یک سری تغییرات عمده در الگوریتم مذکور به منظور بهبود کیفیت و قوی‌بودن خوشه‌ها اعمال شده است.

گام ۱: اولین قدم در الگوریتم پیشنهادی، کنارگذاشتن گره‌های با درجه صفر از روال خوشه‌بندی است. این موضوع به این دلیل است که گره‌های با درجه صفر یعنی گره‌هایی که با هیچ گرهی مجاور نیستند، خود یک جامعه تک‌عضوی جداگانه محسوب می‌شوند و در واقع به دلیل ایزوله بودن آنها معیار خاصی برای انتساب آنها به یک جامعه خاص مبتنی بر لینک‌های بین گره‌ها وجود ندارد. البته این امکان وجود دارد که پس از مشخص‌شدن جامعه‌ها از الگوریتم اصلی، هر یک از گره‌های منفرد و ایزوله یا به صورت تصادفی و یا براساس ملاک‌های دیگری غیر از اتصالات (یال‌ها) به یکی از جوامع استخراج‌شده اضافه شوند.

در این گام همچنین گره‌های با درجه ۱ از مجموعه گره‌هایی که شانس سرخوشه‌شدن خواهند داشت کنار گذاشته می‌شوند. علت این است که همان طور که خواهیم داد الگوریتم پیشنهادی و نیز الگوریتم LBLD [۱] برای تعیین

سرخوشه‌ها بین هر دو گره مجاور، میزان اهمیت آن دو را از نظر مشابهت در همسایه‌هایشان مبنا قرار می‌دهند و بنابراین این موضوع برای گره‌های درجه ۱ که فقط یک گره مجاور دارند منتفی است. البته بدیهی است که ماهیت و دلیل کنارگذاشتن گره‌های درجه ۱ با گره‌های درجه صفر متفاوت است. در واقع گره‌های درجه صفر به کلی از جامعه‌بندی اولیه کنار گذاشته می‌شوند و هر کدام به تنهایی به عنوان یک جامعه تک‌عضوی در نظر گرفته می‌شوند ولی گره‌های درجه ۱ فقط احتمال سرخوشه‌بودن را از دست می‌دهند ولی به عنوان یک عضو عادی از یکی از جوامع شناسایی‌شده گروه‌بندی خواهند شد.

گام ۲: در گام بعدی الگوریتم به ازای هر دو عنصر از شبکه (گراف) میزان شباهت آن دو براساس معیار شباهت جاکارد محاسبه می‌شود. این معیار که در رابطه (۱) بیان شد، یکی از مهم‌ترین معیارهای شباهت در بین رأس‌های گراف است و برای دو رأس i و j از گراف نسبت تعداد رأس‌هایی که همسایه‌های مشترک هر دو رأس i و j هستند به تعداد رأس‌هایی که همسایه حداقل یکی از این دو هستند را نشان می‌دهد.

گام ۳: هر گرهی که بیشترین مشابهت را با همسایه‌هایش داشته باشند به این معنی است که یک زیرگراف با اتصالات زیاد در پیرامون آن وجود دارد و بنابراین این گره، کاندید خوبی برای انتخاب‌شدن به عنوان عنصر مرکزی جامعه یا همان سرخوشه است. به همین دلیل، مشابه الگوریتم LBLD، معیار اهمیت گره (NI)[1] طبق رابطه زیر تعریف می‌شود[۱]:

$$NI_i = \sum_{j \in N_i} Similarity_{ij} \qquad (۲)$$

- این معیار برای هر یک از رأس‌های گراف محاسبه می‌شود و رأس‌ها براساس مقدار NI به صورت نزولی مرتب می‌شوند.
- برای استخراج جوامع، ابتدا همه گره‌های گراف با برچسب صفر برچسب‌گذاری می‌شوند و سپس مراحل زیر برای تمامی گره‌ها (بجز گره‌های درجه صفر و ۱ که قبلاً کنار گذاشته شده بودند) تکرار می‌شود:
- گره i و شبیه‌ترین گره به آن به نام j انتخاب می‌شوند.
- اگر similarity(i,j) ≠ 0 باشد، به هریک از دو گره i و j که NI آن بیشتر است، برچسب جدید انتساب داده می‌شود.
- درغیراین‌صوت (اگر similarity(i,j) = 0 باشد) به این معنی است که گره i نه تنها با گره j بلکه با هر گره دیگر گراف، هیچ همسایه مشترکی ندارد و در نتیجه بالاترین شباهت گره i با بقیه گره‌ها برابر با صفر شده است. در این صورت برچسب همسایه با بالاترین درجه را به هر دو گره انتساب داده می‌شود.

گام ۴: درگام بعدی پنج درصد گره‌های با بیشترین اهمیت انتخاب می‌شود و برچسب آنها به شبیه‌ترین همسایه‌های آنها انتشار (تعمیم) داده می‌شود. اگر تعداد اعضای بعضی از جوامع از حد آستانه معینی کمتر باشد، این جوامع در هم ادغام می‌شوند تا جوامعی با تعداد اعضای مناسب تشکیل شود.

گام ۵: ایده و نوآوری اصلی راهکار پیشنهادی ما در این گام اعمال می‌گردد:

- برای هر یک از جوامع (خوشه‌های) شناسایی‌شده در مرحله قبل مثل c، معیار s_k به صورت زیر تعریف می‌شود:

$$s_k(c) = \frac{\text{number of members in c which have at least k links with other members}}{\text{total number of members in c}} \qquad (۳)$$

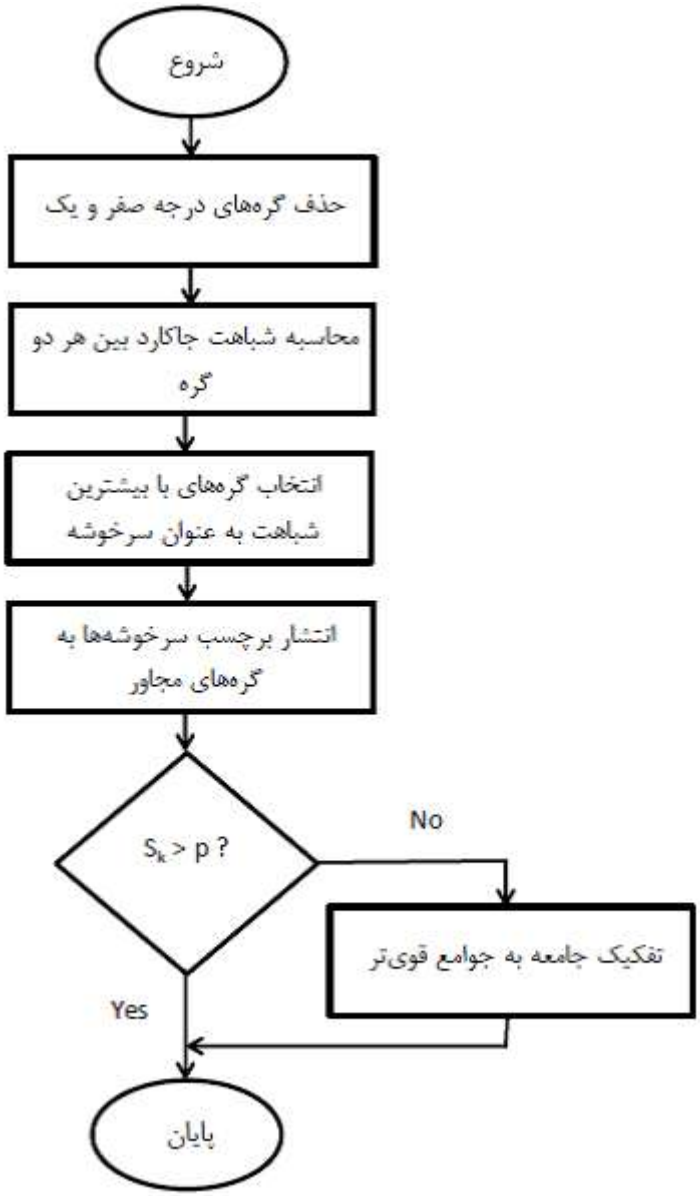
- سپس برای هر جامعه بررسی می‌شود که آیا مقدار s_k آن از یک مقدار حد آستانه p بیشتر است یا نه. اگر بیشتر باشد، جامعه به همین صورت حفظ می‌شود. در غیراین‌صورت جامعه باید به جوامع کوچک‌تر و قوی‌تری تفکیک شود. برای این منظور الگوریتم شناسایی جوامع به صورت بازگشتی روی جامعه مورد نظر اجرا می‌شود. مقدار k به میزان قوی‌بودن مطلوب جامعه بستگی دارند و بدیهی است که حداقل آن ۱ و حداکثر آن n-1 است که n تعداد رأس‌های جامعه (زیرگراف) را نشان می‌دهد.

[1] Node Importance

نکته قابل ذکر در آخرین مرحله الگوریتم پیشنهادی یعنی مرحله شکستن (تفکیک) جوامع ضعیف و تشکیل جوامع قوی‌تر با یک رویکرد بازگشتی صورت می‌گیرد به این ترتیب که مجدداً در این مرحله، الگوریتم انتشار برچسب با استفاده از معیار h-index بهبودیافته برروی جوامعی که نیاز به تفکیک بیشتر دارند، اجرا می‌گردد.

مقدار k و p به میزان قوی‌بودن مطلوب جامعه بستگی دارند. بدیهی است که حداقل مقدار k برابر با ۱ و حداکثر آن n-1 است که n تعداد رأس‌های جامعه (زیرگراف) را نشان می‌دهد. مقدار p نیز درصد اعضایی از جامعه را نشان می‌دهد که حداقل با k عضو دیگر اتصال مستقیم دارند و بنابراین عددی بین صفر تا صددرصد است.

فلوچارت روش پیشنهادی در شکل (۳) نشان داده شده است.



شکل ۳. فلوچارت روش پیشنهادی

۴- ارزیابی روش پیشنهادی

۴-۱- رویکرد ارزیابی و ابزار پیاده‌سازی

برای ارزیابی روش پیشنهادی، این روش پیاده‌سازی و بر روی چند مجموعه‌داده شناخته‌شده در این حوزه اعمال شده است و براساس معیارهای متداول در زمینه شناسایی جوامع با روش ارائه‌شده در مرجع [۱] مقایسه صورت گرفته است.

پیاده‌سازی روش پیشنهادی با استفاده از نرم‌افزار Gephi انجام شده است. این نرم‌افزار یک محیط متن‌باز و رایگان است که به عنوان یکی از متداول‌ترین و کارآمدترین ابزارهای تحلیل و بصری‌سازی گراف‌ها و شبکه‌ها به صورت گسترده استفاده می‌شود و کارهای تحقیقاتی متعددی از جمله [۱۹-۱۷] از آن برای پیاده‌سازی الگوریتم‌های پیشنهادی خود در حوزه مدل‌سازی و تحلیل شبکه‌های اجتماعی استفاده کرده‌اند.

۴-۲- معیارهای ارزیابی

در حوزه شناسایی جوامع، برای ارزیابی و مقایسه کیفیت جوامع (خوشه‌های) استخراج‌شده، معیارهای مختلفی مورد استفاده قرار می‌گیرند که از جمله می‌توان به NMI (اطلاعات متقابل نرمال‌شده[1]) و پیمانگی (ماژولاریتی) اشاره کرد. این دو معیار هر دو نشان می‌دهند که تا چه حد میزان اتصالات در داخل هر جامعه قوی و اتصالات بین جوامع مختلف، ضعیف است.

معیارهای دیگری که برای این منظور استفاده می‌شوند عبارتند از دقت[2]، یادآور یا فراخوانی[3] و F1[4] که روابط مربوط به آنها به صورت زیر است [۱]:

$$\text{Precision} = \frac{|C_D \cap C_R|}{|C_D|} \quad (۴)$$

$$\text{Recall} = \frac{|C_D \cap C_R|}{|C_R|} \quad (۵)$$

$$\text{F} - \text{measure} = \frac{2 * \text{precision} * \text{recall}}{\text{Precision} + \text{recall}} \quad (۶)$$

که در رابطه‌های فوق C_D جامعه‌های شناسایی‌شده توسط الگوریتم مورد بررسی و C_R جامعه‌های واقعی براساس آنچه در مجموعه‌داده مشخص شده است را نشان می‌دهد. در واقع معیار دقت نشان می‌دهد که چه کسری از افراد یک جامعه شناسایی‌شده واقعاً عضو آن جامعه هستند و معیار فراخوانی نشان می‌دهد که چه کسری از افراد یک جامعه واقعی در جامعه شناسایی‌شده انتخاب شده‌اند و معیار F1 نوعی میانگین بین دو معیار قبلی است.

۴-۳- مجموعه‌داده‌های مورد استفاده در ارزیابی

چندین مجموعه داده مطرح برای ارزیابی الگوریتم‌های شناسایی جوامع در کارهای تحقیقاتی مشابه مورد استفاده قرار گرفته است که مهم‌ترین آنها که برای ارزیابی این تحقیق مورد استفاده قرار گرفته است، عبارتند از:

۱- مجموعه‌داده باشگاه کاراته زاخاری: این مجموعه‌داده، یک شبکه اجتماعی است که از یک باشگاه کاراته به مدت سه سال از سال ۱۹۷۰ تا ۱۹۷۲ توسط واین زاخاری موردمطالعه قرار گرفت. این شبکه ۳۴ عضو یک باشگاه کاراته را جذب می‌کند و ۷۸ ارتباط بین جفت عضوهایی که در خارج از باشگاه تعامل دارند، وجود دارد. در طول مطالعه، اختلاف بین مدیر «جان.ای» و مربی «آقای های» (نام مستعار) به وجود آمد که منجر به تقسیم باشگاه به دو بخش شد. نیمی از اعضا در اطراف «آقای های» یک باشگاه جدید تشکیل دادند؛ اعضای بخش دیگر یک مربی جدید پیدا کردند یا از کاراته دست کشیدند. در نهایت بر اساس داده‌های جمع‌آوری‌شده زاخاری، دو جامعه شکل گرفتند.

---

[1] Normalized Mutual Information
[2] Precision
[3] Recall
[4] F-Score /F-measure

۲- مجموعه داده دلفین‌ها: این مجموعه‌داده، یک شبکه اجتماعی بدون جهت از ارتباطات مکرر بین ۶۲ دلفین که با هم بازی می‌کنند، تشکیل شده است. این دلفین‌ها در نیوزیلند زندگی می‌کنند. گروه دلفین‌ها عمدتاً به دو جامعه تقسیم می‌شوند: دلفین‌های نر و دلفین‌های ماده.

۳- مجموعه‌داده فوتبال: این مجموعه‌داده، شبکه‌ای از بازی‌های فوتبال آمریکایی است که بین دانشگاه‌های بخش آیووا برگزار شده است. این شبکه دارای ۱۱۵ بازیکن و ۶۱۳ یال است. یال‌های بین بازیکن‌ها نشان دهنده تعداد پاس‌هایی است که هر بازیکن به بازیکن دیگر می دهد. و به دوازده جامعه تقسیم می‌شوند که شامل مجموعه بازیکن‌های فوتبالی است، که در طول فصل پاییز سال ۲۰۰۰ با هم به رقابت پرداختند.

۴- مجموعه‌داده آمازون: این مجموعه‌داده، شبکه‌ای از کتاب‌های مربوط به سیاست ایالات متحده است که نزدیک به انتخابات ریاست جمهوری ایالات متحده در سال ۲۰۰۴ منتشر شده است. این شبکه دارای ۱۰۵ کتاب و ۴۴۱ یال است. یال‌های بین کتاب‌ها نشان دهنده خرید مشترک مکرر آن کتاب‌ها توسط همان خریداران است. این شبکه توسط والدیس کربس گردآوری شده است و توسط شرکت آمازون به فروش رسید و به سه جامعه تقسیم می‌شوند که شامل سه مجموعه کتاب‌های مرتبط به هم است.

مشخصات این مجموعه‌داده‌ها در جدول (۱) نشان داده شده و توضیح مختصر هریک از آنها در ادامه ارائه شده است.

جدول ۱. مجموعه‌داده‌های درنظرگرفته‌شده برای ارزیابی و مقایسه راهکار پیشنهادی

نام مجموعه داده	تعداد رأس‌ها (افراد)	تعداد یال‌ها (لینک‌ها)	تعداد جوامع (خوشه‌ها)
باشگاه کاراته زاخاری	۳۴	۷۸	۲
دلفین‌ها	۶۲	۱۵۹	۲
فوتبال	۱۱۵	۶۱۳	۱۲
آمازون	۳۳۴۸۶۳	۹۲۵۸۷۲	۷۵۱۴۹

۴-۴- نتایج ارزیابی و مقایسه

اولین آزمایش به منظور تعیین مقدار اطلاعات متقابل نرمال‌شده (NMI) براساس روش پیشنهادی انجام شده است. این بررسی برای چهار مجموعه داده مورد ارزیابی قرار گرفته است. نتایج به‌دست‌آمده در جدول (۲) نشان داده شده و با نتایج روش‌های موجود مقایسه شده است.

جدول ۲. مقادیر NMI حاصل از روش پیشنهادی و مقایسه آن با نتایج روش‌های موجود

نام مجموعه داده	الگوریتم LBLD [۱]	الگوریتم LSMD [۳]	الگوریتم مرجع [۱۶]	روش پیشنهادی
باشگاه کاراته زاخاری	1	1	1	1
دلفین‌ها	1	1	0.98	1
فوتبال	0.91	0.93	0.92	0.95
آمازون	0.97	0.95	0.97	0.98

همان‌گونه که در جدول (۲) مشاهده می‌شود، برای دو مجموعه داده باشگاه کاراته و دلفین‌ها که مجموعه‌هایی با تعداد کم رکورد هستند، الگوریتم پیشنهادی ما مانند دو الگوریتم مورد مقایسه، بهترین نتیجه ممکن یعنی مقدار NMI برابر با 1 را ارائه

می‌کند. در مورد مجموعه‌داده‌های فوتبال و آمازون، مقدار NMI حاصل از راهکار پیشنهادی بیشتر از دو الگوریتم مورد مقایسه است که به معنی انتخاب جوامع قوی‌تر است.

آزمایش بعدی با هدف محاسبه F-measure براساس روش پیشنهادی و مقایسه آن با الگوریتم‌های قبلی است. نتایج این ارزیابی در جدول (۳) نشان داده شده است.

جدول ۳. مقادیر F-measure حاصل از روش پیشنهادی و مقایسه آن با نتایج روش‌های موجود

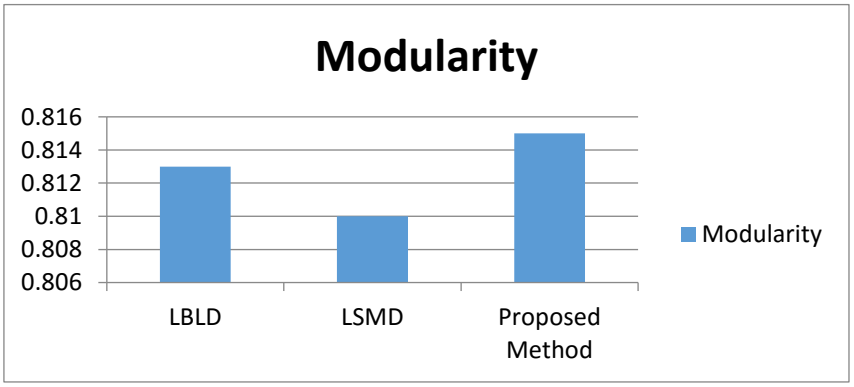
نام مجموعه داده	الگوریتم LBLD [۱]	الگوریتم LSMD [۳]	الگوریتم مرجع [۱۶]	روش پیشنهادی
باشگاه کاراته زاخاری	1	1	1	1
دلفین‌ها	1	1	0.99	1
فوتبال	0.916	0.910	0.914	0.923
آمازون	0.90	0.84	0.87	0.92

مطابق جدول ۳ در مورد معیار F-measure هم نتایج روش پیشنهادی برابر یا قدری بهتر از نتایج دو الگوریتم مورد مقایسه است.

آزمایش بعدی به منظور محاسبه معیار پیمانگی در روش پیشنهادی و مقایسه آن با الگوریتم‌های موجود انجام شد که نتایج آن در شکل ۴ ارائه شده است. لازم به ذکر است که با توجه به این که در مرجع [۱] مقدار این معیار برای مجموعه داده آمازون گزارش شده است بنابراین ارزیابی و مقایسه روش پیشنهادی نیز بر روی همین مجموعه داده انجام گرفته است. شکل ۴ نشان می‌دهد که راهکار پیشنهادی مقدار معیار پیمانگی را به میزان ۲ درصد نسبت به الگوریتم و به میزان ۵ درصد نسبت به الگوریتم LSMD بهبود داده است.

نکته قابل ملاحظه در نتایج حاصل این است که اگرچه اندازه مطلق بهبود مقادیر معیارهای NMI و F-measure در روش پیشنهادی نسبت به سایر روش‌های مورد مقایسه، جزئی به نظر می‌رسد ولی اولاً در این حوزه تحقیقاتی، چون مقدار معیارهای ارزیابی بین صفر و یک است، اغلب بهبودها در این معیارها جزئی است و ثانیاً اگر مقدار نسبی و درصدی بهبود معیارهای ارزیابی مورد توجه قرار گیرد، ملاحظه می‌شود که درصد افزایش مقادیر NMI و F-measure قابل ملاحظه است و نشان می‌دهد که روش پیشنهادی کارایی قابل قبولی در شناسایی جوامع دقیق‌تر از خود نشان داده است.

آخرین آزمایش انجام‌شده برای تعیین مقدار بهینه پارامترهای k و p صورت گرفت که نتایج آن در جدول (۴) ارائه شده است. همان طور که نتایج این جدول نشان می‌دهد بهترین مقادیر F-measure به ازای k=5, p=70% حاصل شده است یعنی شرایطی که در هر جامعه حداقل هفتاد درصد از اعضای جامعه هر یک حداقل با پنج عضو دیگر همان جامعه در ارتباط باشند.



شکل ۴. مقدار معیار پیمانگی حاصل از روش پیشنهادی در مقایسه با الگوریتم‌های LBLD [۱] و LSMD [۳] بر روی مجموعه داده آمازون

جدول ۴. مقادیر F-measure حاصل از روش پیشنهادی به ازای مقادیر مختلف k و p

نام مجموعه داده	k=5, p=60%	k=5, p=70%	k=7, p=60%	k=7, p=70%
باشگاه کاراته زاخاری	0.982	**1**	0.990	0.991
دلفین‌ها	0.976	**1**	0.968	**1**
فوتبال	0.903	**0.923**	0.909	0.916
آمازون	0.914	**0.92**	0.905	0.912

۵- نتیجه‌گیری و کارهای آینده

مسأله استخراج جوامع در شبکه‌های اجتماعی در سال‌های اخیر توجه محققان را به خود جلب کرده است. به دلیل گسترش روزافزون کاربردهای این حوزه و با توجه به اهمیت افزایش دقت استخراج جوامع، در این پایان‌نامه، راهکار جدیدی با بهبود یک الگوریتم موجود در این زمینه موسوم به LBLD [۱] ارائه شد. برای این منظور از معیار شباهت جاکارد برای تعیین میزان مشابهت رأس‌های شبکه از نظر همسایه‌های مشترک استفاده شد. همچنین یک معیار جدید مبتنی بر مفهوم معیار h-index برای افزایش دقت شناسایی جوامع معرفی و استفاده شد. ایده پیشنهادی بر این اصل مبتنی است که برای قوی‌تر بودن یک جامعه شناسایی‌شده لازم است تعداد بیشتری از اعضای جامعه، اتصالات بیشتری با بقیه اعضای جامعه داشته باشند. بنابراین جوامعی که این شرط را برآورده نمی‌کنند در الگوریتم پیشنهادی به جوامع کوچک‌تر و قوی‌تری تفکیک می‌شوند.

در نهایت، راهکار پیشنهادی با پیاده‌سازی و اعمال بر روی چند مجموعه‌داده شناخته‌شده ارزیابی و با کار مرجع [۱] مقایسه گردید. نتایج ارزیابی‌های انجام‌شده نشان‌دهنده بهبود میزان قوی‌بودن جوامع استخراج‌شده نسبت به روش‌های مورد مقایسه از نظر معیارهای NMI، ماژولاریتی و F-measure است. میزان بهبود این سه معیار بر روی مجموعه‌داده‌های مختلف نسبت به الگوریتم‌های مورد مقایسه به طور متوسط برابر با ۳.۲درصد، ۳.۵ درصد و ۲.۹درصد بوده است.

پیشنهادهای مختلفی در راستای ادامه کار این پژوهش و بهبود آن قابل طرح است که از جمله می‌توان به موارد زیر اشاره کرده:

- برای بررسی کیفیت جوامع شناسایی‌شده، علاوه بر ارزیابی میزان اتصالات داخل هر جامعه، میزان اتصالات بین جوامع نیز مورد ارزیابی قرار گیرد زیرا شناسایی جوامع به صورت مطلوب، زمانی حاصل می‌شود که ارتباطات در داخل هر جامعه، بیشینه و بین جوامع مختلف، کمینه باشد.
- پیشنهاد دیگر، استفاده از الگوریتم‌های یادگیری ماشین یا روش‌های فراابتکاری و تکاملی برای تعیین مقدار پارامترهای k و p در روش پیشنهادی است.
- همچنین می‌توان از ترکیب ایده پیشنهادی در این مقاله با الگوریتم‌های خوشه‌بندی کلاسیک و شناخته‌شده به منظور افزایش دقت شناسایی جوامع استفاده کرد.

مراجع

[1] H. Roghani, and A. Bouyer, "A Fast Local Balanced Label Diffusion Algorithm for Community Detection in Social Networks," IEEE Transactions on Knowledge and Data Engineering · January 2022.

[2] K. Berahmand, A. Bouyer, and M. Vasighi, "Community Detection in Complex Networks by Detecting and Expanding Core Nodes Through Extended Local Similarity of Nodes," IEEE Transactions on Computational Social Systems, vol. 5, no. 4, pp. 1021-1033, 2018, doi: 10.1109/TCSS.2018.2879494.

[3] A. Bouyer and H. Roghani, "LSMD: A fast and robust local community detection starting from low degree nodes in social networks," Future Generation Computer Systems, vol. 113, pp. 41-57, 2020/12/01/ 2020, doi: https://doi.org/10.1016/j.future.2020.07.011.

[4] Adamic LA, Glance N, editors. The political blogosphere and the 2004 US election: divided they blog. Proceedings of the 3rd international workshop on Link discovery; 2005.

[5] Z. Sun et al., "Community detection based on the Matthew effect," Knowledge-Based Systems, vol. 205, p. 106256, 2020.

[6] S. Aghaalizadeh, S. T. Afshord, A. Bouyer, and B. Anari, "A three-stage algorithm for local community detection based on the high node importance ranking in social networks," Physica A: Statistical Mechanics and its Applications, vol. 563, p. 125420, 2021.

[7] M. Zarezade, E. Nourani, and A. Bouyer, "Community detection using a new node scoring and synchronous label updating of boundary nodes in social networks," Journal of AI and Data Mining, vol. 8, no. 2, pp. 201-212, 2020.

[8] S. Taheri and A. Bouyer, "Community Detection in Social Networks Using Affinity Propagation with Adaptive Similarity Matrix," Big Data, vol. 8, no. 3, pp. 189-202, 2020.

[9] A. Clauset, M. E. Newman, and C. Moore, “Finding community structure in very large networks,” Physical Review E, vol. 70, no. 6, p. 066111, 2004.

[10] Yang Z, Algesheimer R, Tessone CJ. A comparative analysis of community detection algorithms on artificial networks. Scientific reports. 2016;6(1):30750.

[11] F. D. Zarandi and M. K. Rafsanjani, “Community detection in complex networks using structural similarity,” Physica A: Statistical Mechanics and its Applications, vol. 503, pp. 882–891, 2018.

[12] M’barek MB, Hmida SB, Borgi A, Rukoz M. GA-PPI-Net Approach vs Analytical Approaches for Community Detection in PPI Networks. Procedia Computer Science. 2021;192:903-12.

[13] Patil SV, Kulkarni DB, editors. Graph partitioning using heuristic Kernighan-Lin algorithm for parallel computing. Next Generation Information Processing System: Proceedings of ICCET 2020, Volume 2; 2021: Springer.

[14] Gharehchopogh, F.S., 2023. An improved Harris Hawks optimization algorithm with multi-strategy for community detection in social network. Journal of Bionic Engineering, 20(3), pp.1175-1197.

[15] Hevey D. Network analysis: a brief overview and tutorial. Health psychology and behavioral medicine. 2018;6(1):301-28.

[16] V. A. Traag, L. Waltman, and N. J. van Eck, "From Louvain to Leiden: guaranteeing well-connected communities," Scientific reports, vol. 9, no. 1, pp. 1-12, 2019.

[17] Hernández-García, Á., Cuenca-Enrique, C., Traxler, A., López-Pernas, S., Conde-González, M. Á., & Saqr, M. (2024). Community detection in learning networks using R. In Learning analytics methods and tutorials: a practical guide using R (pp. 519-540). Cham: Springer Nature Switzerland.

[18 Yilmaz, C. B., Dagdeviren, Z. A., & Unalir, M. O. (2024, September). Identifying Social Connections: An Empirically-Driven Evaluation of Community Detection Methods on Twitter. In 2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP) (pp. 1-8). IEEE.

[19] Rostami, M., Oussalah, M., Berahmand, K., & Farrahi, V. (2023). Community detection algorithms in healthcare applications: A systematic review. *IEEE Access*, *11*, 30247-30272.

[20] Su, S., Guan, J., Chen, B., & Huang, X. (2023). Nonnegative matrix factorization based on node centrality for community detection. *ACM Transactions on Knowledge Discovery from Data*, *17*(6), 1-21.

Increasing Community Detection Precision in Social Networks using Improved Label Diffusion Approach

Nafiseh Afkhami, MSc [1] | Nazbanoo Farzaneh, Assistant Professor [2] | Hassan Shakeri, Associate Professor[3] *

[1] Deprtment of Computer Engineering, Imam Reza International University, Mashhad, Iran email: nafiseafkhami64@gmail.com

[2] Deprtment of Computer Engineering, Imam Reza International University, Mashhad, Iran, nazbanou.farzaneh@imamreza.ac.ir

[3] Deprtment of Computer Engineering, Ma.C., Islamic Azad University, Mashhad, Iran, email: hassanshakeri@iau.ac.ir

Correspondence
Hassan Shakeri, Associate Professor, Deprtment of Computer Engineering, Ma.C., Islamic Azad University, Mashhad, Iran, email: hassanshakeri@iau.ac.ir

Abstract

Detecting communities in large networks is an important challenge in social network analysis, and providing an algorithm with optimal precision and efficiency for extracting communities is very important. There are different approaches to identify communities in social networks, including methods based on classical clustering, algorithms based on criteria of feature similarity, methods based on finding subgraphs with a lot of internal communication, as well as the label propagation approach. In the label propagation approach, first, the most important nodes of the network are determined based on a series of importance and centrality criteria, and different community labels are assigned to them. Then the label of each of these nodes is propagated to its neighbor nodes. The aim of this research is to improve the so-called LBLD community detection algorithm. This algorithm first determines five percent of the most important network nodes based on a similarity criterion. Then, with a balanced approach, communities are developed both from the center and the borders, and finally a phase of integration is implemented so that small communities are combined with each other and form desirable communities. Our proposed idea uses a measure inspired by the concept of h-index to improve the precision of community detection. In such a way those subgraphs are identified as communities that have at least p percent of nodes with at least degree k. The precision and efficiency of the proposed solution have been evaluated by applying it to some well-known datasets in this field and it shows a significant improvement compared to existing similar methods.

Keywords: community detection, social network modeling, label propagation algorithm, clustering.