نشریه علمی-تخصصی
پژوهش‌های نوین علوم مهندسی
ISSN: 2538-3809

سال یازدهم، شماره ۳ (پیاپی ۵۷)، تابستان ۱۴۰۵

ارائه روش‌های مبتنی بر اتوماتای یادگیر جهت شناسایی اجتماعات در شبکه‌های اجتماعی آنلاین

سحر بافکار[۱]

تاریخ دریافت: ۱۴۰۵/۰۲/۲۶
تاریخ پذیرش: ۱۴۰۵/۰۴/۳۰
کد مقاله: ۳۳۴۲۷

چکـیده

شناسایی اجتماعات در شبکه‌های اجتماعی آنلاین، یکی از مسائل بنیادین در تحلیل شبکه‌های پیچیده است که کاربردهای گسترده‌ای در بازاریابی هدفمند، سامانه‌های توصیه‌گر و امنیت سایبری دارد. علی‌رغم پیشرفت‌های قابل‌توجه، الگوریتم‌های موجود معمولاً با محدودیت‌هایی نظیر نیاز به دانستن تعداد اجتماعات، پیچیدگی محاسباتی بالا) به‌ویژه در روش‌های طیفی با مرتبه($O(n^3)$) ، یا عدم کارایی مناسب بر روی گراف‌های وزن‌دار و پویا مواجه هستند. در این پژوهش، یک روش نوین برای شناسایی اجتماعات در گراف‌های ایستا (بدون وزن و وزن‌دار) ارائه می‌شود که مبتنی بر ترکیب اتوماتای یادگیر توزیع‌شده و یک راهبرد جدید خوشه‌بندی مبتنی بر اشتراک همسایگی است. در الگوریتم پیشنهادی، به هر رأس گراف یک اتوماتای یادگیر تخصیص داده می‌شود. الگوریتم در سه مرحله اصلی شامل پیمایش تطبیقی گراف، ارزیابی مسیر طی‌شده با استفاده از تابع کیفیت ماژولاریتی (Q) و به‌روزرسانی پاداش–جریمه، و در نهایت خوشه‌بندی رئوس بر اساس اشتراک همسایگی، عمل می‌کند. دو نسخه از الگوریتم به ترتیب برای گراف‌های بدون وزن (CDUDLA) و وزن‌دار (CDWDLA)پیاده‌سازی شده‌اند. ارزیابی روی هفت دادگان استاندارد شاملZachary Karate Club ، Dolphins، American Football، و دادگان مصنوعی LFR با مقیاس تا ۱۰٬۰۰۰ گره انجام شده است. نتایج نشان می‌دهد که الگوریتم پیشنهادی در حالت بهینه (a=0.5, b=0.002) به ماژولاریتی ۴۱۲۵/۰ برای داده کاراته، ۵۳۴۹/۰ برای دلفین‌ها و ۶۳۱۵/۰ برای فوتبال دست می‌یابد که در مواردی بهتر از روش‌های مطرحی چونGN ، Blondelو LDF است. همچنین، افزایش نسبت پاداش به جریمه (a/b) باعث کاهش معنادار زمان اجرا و بهبود خلوص و NMI می‌شود. تحلیل حساسیت نشان می‌دهد الگوریتم تا نرخ نمونه‌برداری ۵۰٪، همبستگی بالایی با ساختار اصلی اجتماعات دارد. نوآوری اصلی در استفاده همزمان از یادگیری تقویتی توزیع‌شده و معیار اشتراک همسایگی برای خوشه‌بندی بدون نیاز به تعداد اجتماعات از پیش تعیین‌شده است.

واژگـان کلـیدی: شبکه‌های اجتماعی آنلاین، خوشه‌بندی، اتوماتای یادگیر توزیع شده، نمونه‌برداری

۱ - مهندسی کامپیوتر گرایش هوش مصنوعی، دانشگاه آزاد اسلامی واحد قزوین. ایران
s.bafkar@gmail.com

۱- مقدمه

شبکه‌های اجتماعی آنلاین به عنوان یکی از مهم‌ترین پدیده‌های عصر دیجیتال، نقشی کلیدی در شکل‌گیری تعاملات انسانی، انتشار اطلاعات و ایجاد جوامع مجازی ایفا می‌کنند. این شبکه‌ها را می‌توان به صورت گراف‌هایی مدل‌سازی کرد که در آن‌ها نودها معرف افراد یا موجودیت‌ها و یال‌ها نشان‌دهنده روابط میان آن‌ها هستند .(Fortunato, 2010)بسیاری از پژوهش‌های اولیه در حوزه نظریه گراف، عمدتاً بر روی گراف‌های ایستا متمرکز بوده‌اند، اما در کاربردهای واقعی، شبکه‌های اجتماعی ذاتاً پویا هستند و ساختارشان در طول زمان تغییر می‌کند فیسبوک، توییتر و سایر پلتفرم‌های مشابه، نمونه‌هایی از این شبکه‌های پویا هستند که یال‌ها و نودهای جدید به آن‌ها افزوده یا از آن‌ها حذف می‌شوند-(Tantipathananandh, Tanya. Berger .(Wolf,2011یکی از مسائل اساسی و چالش‌برانگیز در تحلیل این شبکه‌ها، شناسایی اجتماعات است. اجتماع به گروهی از نودها گفته می‌شود که نسبت به سایر نقاط شبکه، ارتباطات درونی محکم‌تر و تراکم بیشتری دارند در شبکه‌های اجتماعی آنلاین، این اجتماعات می‌توانند نشان‌دهنده گروه‌هایی با علایق مشترک، روابط کاری، دوستی‌های نزدیک یا هم‌محلی باشند (Fortunato,2010) کشف این ساختارها، کاربردهای گسترده‌ای در زمینه‌هایی مانند بازاریابی هدفمند، توصیه‌گرهای اجتماعی، شناسایی گروه‌های تأثیرگذار و حتی امنیت سایبری دارد.روش‌های متنوعی برای شناسایی اجتماعات ارائه شده است که از آن جمله می‌توان به روش‌های پارتیشن‌بندی گراف ، خوشه‌بندی سلسله مراتبی، خوشه‌بندی طیفی ، روش‌های تقسیم‌کننده مانند الگوریتم گیروان–نیومن (GN) و روش‌های مبتنی بر بهینه‌سازی ماژولاریتی نظیر الگوریتم حریصانه بلوندل و شبیه‌سازی تبرید اشاره کرد.(Fortunato ,2010) با وجود پیشرفت‌های قابل توجه، بسیاری از این الگوریتم‌ها با محدودیت‌هایی مواجه هستند. برای مثال، روش‌های پارتیشن‌بندی نیازمند دانستن تعداد اجتماعات به عنوان ورودی هستند روش‌های طیفی برای گراف‌های بزرگ با پیچیدگی محاسباتی $O(n^3)$ عملاً غیرقابل اجرا هستند و برخی روش‌های حریصانه نیز از دقت کافی برخوردار نیستند.(Harvey, Kleinberg ,Lehman ,2006) همچنین، گراف‌های واقعی اغلب وزن‌دار، بدون جهت یا جهت‌دار هستند و بسیاری از الگوریتم‌ها تنها برای گراف‌های ساده و بدون وزن طراحی شده‌اند.از سوی دیگر، شبکه‌های اجتماعی آنلاین به دلیل مقیاس بسیار بزرگ و همچنین تمایل نداشتن به اشتراک‌گذاری کامل داده‌ها، نیازمند روش‌هایی هستند که بتوانند با استفاده از نمونه‌برداری از گراف اصلی، به شناسایی اجتماعات بپردازند (ریحانیان، علیزاده، مینایی، ۱۳۹۲).این مسئله اهمیت توسعه الگوریتم‌هایی را که هم از دقت بالا و هم از پیچیدگی محاسباتی مناسب برخوردار باشند، دوچندان می‌کند.در این راستا، استفاده از اتوماتای یادگیر توزیع‌شده رویکردی نوین و امیدوارکننده است.(شیخ احمدی، دالوند، کشمیری، ۱۴۰۱) اتوماتای یادگیر با دریافت سیگنال‌های تقویتی از محیط، به تدریج بهترین عمل را از میان مجموعه اعمال خود یاد می‌گیرد. با توزیع این اتوماتاها روی گره‌های گراف، می‌توان یک فرآیند یادگیری تطبیقی برای کشف ساختار پنهان اجتماعات ایجاد کرد.(Xiang, 2012)با توجه به مبانی فوق، سؤال پژوهش اصلی این پایان‌نامه به این صورت تدوین می‌شود:

آیا ترکیب اتوماتای یادگیر توزیع‌شده با یک روش جدید خوشه‌بندی مبتنی بر اشتراک همسایگی، می‌تواند به الگوریتمی کارآمد برای شناسایی اجتماعات در گراف‌های ایستا (وزن‌دار و بدون وزن) و پویا منجر شود که ضمن داشتن دقت قابل مقایسه یا بهتر از روش‌های موجود، از سرعت اجرای بالاتری برخوردار بوده و قابلیت اعمال بر روی گراف‌های بزرگ با استفاده از نمونه‌برداری را نیز دارا باشد؟ پاسخ به این سؤال، هدف اصلی این پژوهش را تشکیل می‌دهد و در فصول آتی، الگوریتم پیشنهادی CDUDLA) برای گراف‌های بدون وزن و CDWDLA برای گراف‌های وزن‌دار (معرفی، پیاده‌سازی و با استفاده از معیارهایی همچون خلوص، NMIو ماژولاریتی مورد ارزیابی قرار خواهد گرفت.

۲- روش پژوهش

اجرای پژوهش به صورت شبه‌تجربی و گذشته‌نگر طراحی شده است، زیرا الگوریتم پیشنهادی بر روی داده‌های موجود و واقعی ارزیابی می‌شود بدون دخالت مستقیم در فرآیند تولید داده‌ها.جامعه آماری شامل کلیه گراف‌های استاندارد حوزه شناسایی اجتماعات است. نمونه مورد بررسی شامل هفت دادگان معتبر شامل باشگاه کاراته زاخاری (۳۴ نود، ۷۸ یال) شبکه اجتماعی دلفین‌ها (۶۲ نود، ۱۵۹ یال) کالج فوتبال آمریکا (۱۱۵ نود، ۶۱۳ یال) دادگان میزِ شام (۲۶ نود، ۴۴ یال) و سه دادگان مصنوعی LFRبا مقیاس‌های ۵۰، ۵۰۰، ۱۰۰۰ و ۱۰۰۰۰ نود می‌باشند. روش تجزیه‌وتحلیل: الگوریتم پیشنهادی CDUDLA برای گراف‌های بدون وزن و CDWDLA برای گراف‌های وزن‌دار در محیط نرم‌افزاری پیاده‌سازی و روی هر دادگان اجرا شد. برای هر آزمایش، پارامترهای یادگیری) پاداش a و جریمه (b در مقادیر مختلف (۰۰۱/۰، ۰۰۵/۰، ۰۱/۰، ۲/۰، ۵/۰) تنظیم گردید. ارزیابی نتایج با استفاده از سه معیار اصلی خلوص اطلاعات متقابل نرمال‌شده (NMI) و ماژولاریتی[1] انجام شد. همچنین زمان

[1] *Modularity*

اجرای الگوریتم بر حسب ثانیه اندازه‌گیری گردید. نتایج الگوریتم پیشنهادی با ۱۰ روش مطرح شاملGN ، بلوندل، LDF، OPT، CNMو WT مقایسه شد. کلیه آزمایش‌ها در شرایط یکسان و با حداقل ۳۰ بار تکرار برای اطمینان از پایداری نتایج انجام شده است.

۳- مبانی نظری

۳-۱- شناسایی اجتماعات گراف

شناسایی اجتماعات گراف یک موضوع محبوب در علم کامپیوتر به حساب می‌آید. برای مثال در محاسبات موازی[1]، دانستن بهترین راه برای تخصیص وظایف به پردازنده به طوری‌که ارتباط بین آن‌ها را به حداقل برساند و عملکرد سریع محاسبه را به دنبال داشته باشد، بسیار مهم می‌باشد. این مسئله را می‌توان با تقسیم خوشه‌های کامپیوتری به دو گروه با تعداد پردازنده مساوی، به طوری‌که تعداد اتصالات فیزیکی بین پردازنده‌های گروه‌های مختلف حداقل شود انجام داد. (Yarahmadi , Akbari.Zandevakili,2012). تعریف ریاضی این مسئله پارتیشن‌بندی گراف نام دارد، که جزو دسته روش‌های سنتی می‌باشد. الگوریتم‌های پارتیشن‌بندی گراف برای شناسایی اجتماعات خوب نیست، زیرا لازم است که به عنوان ورودی تعداد گروه‌ها و در برخی موارد حتی سایز هر گروه را بدانیم(Harvey, Kleinberg, Lehman, 2006).

۳-۲- خوشه‌بندی سلسله مراتبی

از دیگر روش‌های این دسته می‌توان به خوشه‌بندی سلسله مراتبی[2]، خوشه‌بندی پارتیشنی[3] و خوشه‌بندی طیفی[4] اشاره کرد. شبکه‌های اجتماعی، اغلب ممکن است یک ساختار سلسله‌مراتبی داشته باشند. در چنین مواردی، ممکن است از الگوریتم‌های خوشه‌بندی سلسله مراتبی استفاده شود. تکنیک‌های خوشه‌بندی، ساختار چند سطحی گراف شبکه را نشان می‌دهند. خوشه‌بندی سلسله مراتبی در تجزیه و تحلیل شبکه‌های اجتماعی، زیست‌شناسی، مهندسی، بازاریابی، و غیره بسیار معمول است. (Dinh, Thang, and My Thai,2013)در این روش نیازی به دانستن تعداد و سایز هر خوشه نداریم ولی در عین حال رئوس یک جامعه ممکن است که درست طبقه نشوند و یا به عبارت دیگر عدم مقیاس‌بندی خوب اصلی‌ترین ضعف این روش می‌باشد. روش خوشه‌بندی پارتیشنی یکی دیگر از روش‌های محبوب برای پیدا کردن خوشه‌ها در مجموعه‌ای از نقاط داده‌ها می‌باشد. در اینجا، تعداد خوشه‌ها از قبل تخصیص داده شده است.(Duan, Dongsheng, Yuhua, Ruixuan, Zhengding,2012) این نقاط در یک فضای متریک تعبیه شده‌اند، به طوری‌که هر راس یک نقطه است و یک معیار فاصله بین جفت نقاط در فضا تعریف شده است. هدف جدا کردن نقاط درk خوشه می‌باشد، به طوری‌که تابع هزینه حداکثر / حداقل شود. تابع هزینه بر اساس فاصله بین نقاط و مرکز ثقل می‌باشد به عبارت دیگر موقعیت مناسب تعریف شده در فضا است. مشکل این روش همانند مشکل روش پارتیشن‌بندی سلسله مراتبی است یعنی باید در ابتدا تعداد کلاسترها مشخص باشد و الگوریتم قادر به مشخص کردن آن نمی‌باشد. (Xia, ZhengYou, Zhan Bu,2012) روش خوشه‌بندی طیفی شامل همه روش‌ها و تکنیک‌هایی است که مجموعه را به خوشه‌هایی با استفاده از بردارهای ویژه ماتریس تقسیم می‌کند. خوشه‌بندی طیفی شامل تبدیل مجموعه اولیه از اشیاء به مجموعه‌ای از نقاط در فضا می‌باشدکه مختصات آن عناصر بردارهای ویژه هستند سپس مجموعه نقاط با استفاده از تکنیک‌های استاندارد خوشه‌بندی می‌شوند. خوشه‌بندی طیفی نیاز به محاسبه اولین k بردار ویژه از ماتریس لاپلاس است. اگر گراف ورودی بزرگ باشد، محاسبات دقیق بردارهای ویژه غیر ممکن خواهد بود به طوری‌که نیاز به زمان $O(n^3)$ خواهد داشت. دسته دیگر روش‌های تقسیم کننده[5] هستند. یک راه ساده برای شناسایی جوامع در یک گراف تشخیص و حذف لبه‌هایی است که رئوس جوامع مختلف را به هم وصل کرده‌اند، به طوری‌که خوشه‌ها را از یکدیگر جدا کند. این روش پیشرفت مفهومی قابل توجهی نسبت به روش‌های سنتی ندارد. (Delvenne, Yaliraki, M.Barahona, 2010)تفاوت اصلی این روش با خوشه‌بندی سلسله مراتبی این است که در اینجا لبه‌های درون خوشه به جای لبه‌های بین جفت از رئوس با شباهت کم حذف می‌شوند. در برخی از موارد ممکن است، رئوس (با تمام لبه‌های مجاور خود) یا کل زیرگراف به جای لبه‌های تکی حذف شوند. این روش برای نشان‌دادن پارتیشن‌های به دست آمده با استفاده از دندروگرام مرسوم است. روش‌های مبتنی بر پیمانه‌ای بودن[6] که شامل تکنیک‌های حریصانه[7] و آب‌دهی

[1] *Parallel Computing*
[2] *Hierarchical clustering*
[3] *Partitional clustering*
[4] *Spectral clustering*
[5] *Divisive algorithms*
[6] *Modularity-based methods*
[7] *Greedy*

فولاد[1] می‌باشد، یکی دیگر از روش‌های شناسایی اجتماعات می‌باشند. در این دسته، پیمانه‌ای بودن[2] یکی از پراستفاده‌ترین و معروف‌ترین توابع کیفیت است. اگر مقدار این تابع زیاد باشد نمایانگر پارتیشن خوب است. بنابراین، پارتیشن مربوطه در گراف داده شده باید بهترین و یا حداقل یک نمونه بسیار خوب باشد. این انگیزه اصلی برای اکثر روش‌های تشخیص اجتماعات می‌باشد. با وجود پیشرفت‌ها و اصلاحات، دقت روش حریصانه، در مقایسه با روش‌های دیگر کافی نیست. روش آب‌دهی فولاد، یک روش احتمالاتی برای بهینه‌سازی در زمینه‌های مختلف است. این روش به طور بالقوه می‌تواند بسیار نزدیک به modularity maximum باشد، اما روش کندی است. پیچیدگی واقعی این روش را نمی‌توان تخمین زد زیرا به شدت به پارامترهای انتخاب شده برای بهینه سازی (دمای اولیه، عامل خنک کننده) بستگی دارد. (Dinh, Thang, My Thai,2013)این روش برای گراف‌های کوچک، تا حدود ۱۰۴ راس استفاده می‌شود. در الگوریتم‌های طیفی[3] ، خواص طیفی ماتریس گراف، غالبا برای پیدا کردن پارتیشن مورد استفاده قرار می‌گیرد. بیشتر روش‌های طیفی در علوم کامپیوتر معرفی شده‌اند و به طور کلی بر روی خوشه‌بندی داده تمرکز می‌کنند. بردارهای ویژه ماتریس انتقال می‌توانند برای استخراج اطلاعات مفید در مورد ساختار جامعه مورد استفاده قرار بگیرند. محدودیت اصلی این روش این است که فرض می‌شود جوامع به صورت حس قوی تعریف شده‌اند که بیش از حد محدود است(McDaid, Aaron ,Derek Greene, Neil Hurley,2011).

۴- الگوریتم پیشنهادی

الگوریتم جهت شناسایی اجتماعات بر روی گراف‌های ساده و وزن‌دار پیشنهاد شده است. در این الگوریتم، ورودی به صورت گرافی از مجموعه رئوس و یال‌های شبکه G=(V,E) می‌باشد و خروجی، اجتماعات به‌دست آمده توسط الگوریتم است. در ابتدا شبکه‌ای از اتوماتاهای یادگیر متناظر با گراف ورودی ایجاد می‌شود، به طوری‌که بر روی هر نود یک اتوماتا قرار می‌گیرد. برای اقدام هر اتوماتا انتخاب یکی از همسایگان در نظر گرفته می‌شود و سیگنال تقویتی در حین اجرای الگوریتم توسط Q (فرمول شماره ۱) به‌دست می‌آید. بردار احتمال اعمال هر اتوماتا در ابتدا به‌طور مساوی معادل با D (فرمول شماره ۳) در نظر گرفته می‌شود. الگوریتم پیشنهادی از سه بخش تشکیل شده است:

۱- پیمایش گراف ورودی توسط اتوماتای یادگیر توزیع شده

۲- ارزیابی مسیر طی شده توسط اتوماتا

۳- خوشه‌بندی نودهای گراف ورودی

۴-۱- پیمایش گراف ورودی توسط اتوماتای یادگیر توزیع شده

پس از مقداردهی اولیه تا رسیدن به شرایط توقف مراحل زیر به ترتیب انجام می‌شوند:

یک نود به صورت تصادفی انتخاب می‌شود.

اتوماتای نظیر آن فعال شده و از بین انتخاب‌های خود یک نود دیگر را انتخاب می‌کند.

اتوماتای متناظر با نود جدید فعال شده و اتوماتای متناظر با نود قبلی غیر فعال می‌گردد.

هر نودی که انتخاب شد از مجموعه نودهای گراف حذف شده تا به صورت تکراری انتخاب نشود.

شرایط توقف:

عدم توانایی اتوماتا در انتخاب نود جدید.

رسیدن به حداکثر تعداد تکرار.

در پایان این مرحله یک مسیر طی شده به صورت کاملا تصادفی[4] خواهیم داشت.

۴-۲- ارزیابی مسیر طی شده توسط اتوماتا

گام بعدی ارزیابی مسیر طی شده توسط اتوماتا و همچنین به‌روزرسانی بردار احتمال می‌باشد، هر نود انتخاب شده توسط اتوماتا، منجر به تغییراتی در مقدار اولیه Q=0 خواهد شد که این تغییرات با توجه به فرمول(۲-۱) محاسبه و بررسی می‌شود.

با توجه به شرایط زیر یکی از تصمیمات عنوان شده اتخاذ خواهد شد:

- اگر نود انتخاب شده منجر به افزایش Q شده باشد بردار احتمال متناظر با آن اقدام به اندازه a تشویق می‌گردد و اعمال تغییرات در Q با انتخاب این نود، فرمول (۲-۴) انجام می‌شود.

[1] *Simulated Annealing*
[2] *Modularity*
[3] *Spectral Algorithms*
[4] *Random Walk*

- اگر نود انتخاب شده منجر به کاهش و یا عدم تغییر در مقدار Q شده باشد بردار احتمال متناظر با آن اقدام به اندازه b جریمه می‌گردد.

پس از به‌روزرسانی بردار احتمال اتوماتا، مجددا این مرحله تکرار می‌شود تا به حداکثر تکرار برسیم. در شروع هر مرحله از تکرار الگوریتم، گراف به صورت کامل به الگوریتم وارد می‌شود، به این معنا که نودهای حذف شده از مجموعه نودهای گراف مجددا به گراف اضافه شده و الگوریتم اجرا می‌شود.

۴-۳- خوشه‌بندی نودهای گراف ورودی

در این بخش باید خوشه‌بندی بر روی نودها بر اساس مسیر پیمایش شده توسط اتوماتا در بخش قبل صورت پذیرد. این خوشه‌بندی بر اساس معیار اشتراک همسایگی نودها انجام خواهد شد. تا زمانی‌که تمامی نودها خوشه‌بندی شوند، برای هر دو نود متوالی انتخاب شده توسط اتوماتا، مراحل زیر انجام می‌گیرد:

همسایه‌های تک‌تک نودها به‌دست می‌آید.

برای نودهای موجود در یک مسیر، دوبه‌دو اشتراک همسایه‌هایشان محاسبه می‌گردد.

در صورتی‌که این اشتراک تهی باشد هر کدام از آن نودها را در یک خوشه منحصر به فرد در نظر می‌گیریم.

در صورتی که این اشتراک غیرتهی باشد هر دو نود را در یک خوشه یکسان در نظر می‌گیریم.

در صورتی که نودهایی وجود داشته باشد، به طوری‌که یا توسط اتوماتا انتخاب نشده باشند و یا حالت جزیره‌ای داشته باشند مراحل زیر انجام می‌گیرد:

هر یک از آن نودها را به ترتیب به تک تک خوشه می‌دهیم و Q جدید هر خوشه را محاسبه می‌کنیم .

اگر با اضافه شدن نود جدید Q خوشه بیشترین بهبود را داشته باشد، می‌توان نتیجه گرفت که نود جدید درست خوشه‌بندی شده است.

در غیر این صورت با خوشه‌های دیگر این کار را تکرار می‌کنیم.

مراحل فوق تا رسیدن به نتیجه مطلوب و نهایی تکرار می‌شود.

$$Q = \sum_i (e_{ii} - {a_i}^2)$$ فرمول (۱-)

که در آن a_i نشان‌دهنده درجه هر راس می‌باشد که با فرمول زیر بدست می‌آید:

$$a_i = \sum_j e_{ij}$$ فرمول (۲-)

$$D = \frac{1}{a_i}$$ فرمول (۳-)

در صورتی‌که ادغام دو نود منجر به افزایش Q شود، تغییرات Q با استفاده از فرمول زیر محاسبه شده و اعمال می‌گردد:

$$\Delta Q = 2(e_{ij} - a_i a_j)$$ فرمول (۴-)

در نهایت میزان مطلوبیت خوشه‌ها با استفاده از تابع ماژولاریتی به نام mod _ metric به‌دست می‌آید.

۵- توضیح دادگان آزمایشی

در این فصل ما از دادگان آزمایشی و واقعی مانند : باشگاه کاراته، شبکه اجتماعی دلفین‌ها، کالج فوتبال، میز شام و همچنین یک سری دادگان LFR که پایگاه‌داده اول شامل۵۰ نود و پایگاه داده دوم ۵۰۰ نود و پایگاه داده سوم ۱۰۰۰ نود می‌باشد استفاده کرده‌ایم. پایگاه داده‌های مذکور در مرجع قابل دسترس می‌باشند.به منظور اجرای الگوریتم، فایل حاوی ماتریس گراف ورودی در برنامه وارد[1] شده و پس از تنظیمات اولیه، الگوریتم را اجرا می‌کنیم.

جدول -۱- دادگان آزمایشی استفاده شده در الگوریتم پیشنهادی اول

Dataset Name	Number of Nodes	Number of Edges	Number of Communities	Status
Zachary Karate Club	34	78	2	Unweighted-Undirected
Dolphin Social Network	62	159	3	Unweighted-Undirected
American Football College	115	613	12	Weighted-Undirected
Dining Table	26	44	3	Weighted-

[1] *Import*

				Undirected
LFR	50	112	3	Unweighted-Undirected
LFR	500	4379	5	Unweighted-Undirected
LFR	1000	8234	113	Unweighted-Undirected

۶- آزمایشها

در این بخش، به‌منظور ارزیابی کارایی الگوریتم‌های پیشنهادی CDUDLA برای گراف‌های بدون وزن و DWDLA برای گراف‌های وزن‌دار، مجموعه‌ای از آزمایش‌های تجربی بر روی دادگان استاندارد و مصنوعی (LFR) انجام شده است. برای سنجش کیفیت ساختار اجتماعاتِ کشف‌شده و همچنین سرعت اجرای الگوریتم، از معیارهای ارزیابی استانداردی شامل خلوص (Purity)، اطلاعات متقابل نرمال‌شده (NMI) و ماژولاریتی (Modularity) استفاده گردیده است. تمامی نتایج حاصل با عملکرد الگوریتم‌های شناخته‌شده‌ای نظیر *GN*، *بلوندل*، *LDF*و *OPT*مقایسه شده‌اند تا توانمندی رویکرد پیشنهادی در مواجهه با چالش‌های شناسایی اجتماع در شبکه‌های اجتماعی نمایان گردد.

۶-۱- بررسی معیار خلوص[1]

هدف: ارزیابی نتایج خوشه‌های به‌دست آمده توسط الگوریتم پیشنهادی بر اساس معیار خلوص.

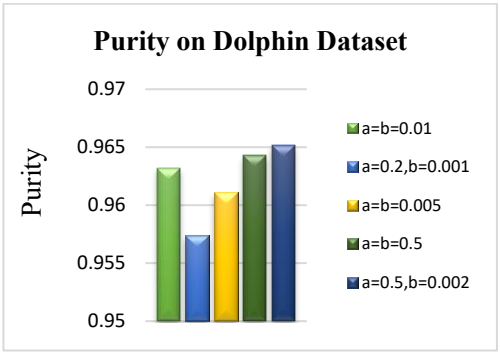
معیار خلوص، میزان منحصربفرد بودن خوشه‌ها را نشان می‌دهد، که عددی بین صفر و یک است. در واقع این معیار نشان می‌دهد که آیا خوشه‌ها با یکدیگر اشتراک دارند یا خیر که هر چه این عدد به یک نزدیک‌تر باشد، یعنی خوشه‌بندی صورت گرفته خالص‌تر است و خوشه‌ها اشتراک کمتری با یکدیگر دارند. این معیار با توجه به فرمول زیر محاسبه می‌گردد:

$$\mathrm{Purity}(\Omega,\mathbb{C}) = \frac{1}{N}\sum_k \max_j \left|\omega_k \cap c_j\right| \qquad \text{فرمول (۲-۵)}$$

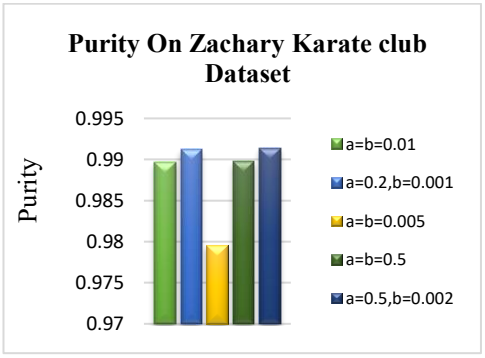
که در آن :

$\mathbf{\Omega} = \{\omega_1, \omega_2, \dots, \omega_k\}$ مجموعه خوشه‌ها و $\mathbb{C} = \{c_1, c_2, \dots, c_j\}$ مجموعه کلاس‌ها می‌باشد.

ما الگوریتم را برروی دادگان آزمایشی مندرج در جدول ۲-۱ آزمایش کردیم و معیار خلوص بدست آمده در هر پایگاه داده با توجه به پارامترهای تنظیم شده به شرح شکل‌های زیر می‌باشد.

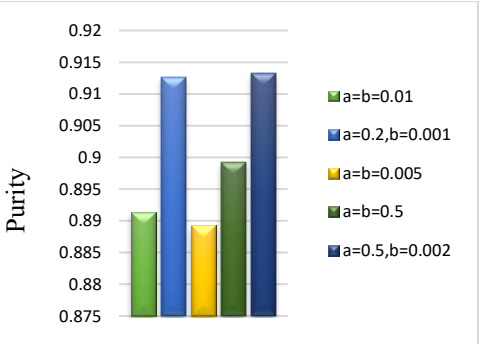


شکل ۱- بررسی معیار خلوص در پایگاه داده دلفین‌ها

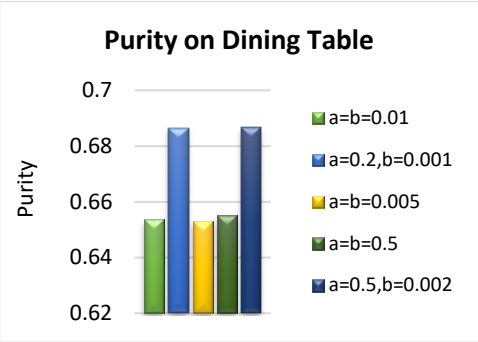


شکل ۲- بررسی معیار خلوص در پایگاه داده باشگاه کاراته

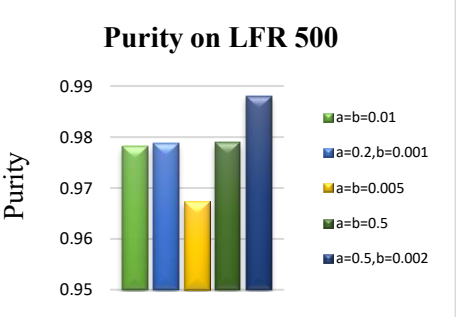
[1] Purity



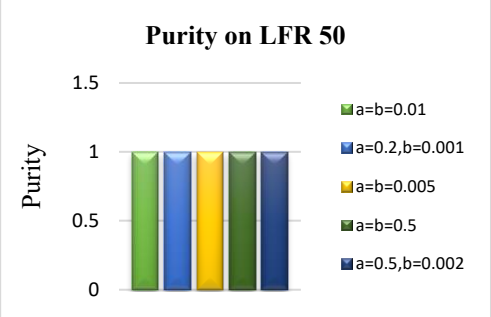
شکل ۵- بررسی معیار خلوص در پایگاه داده کالج فوتبال



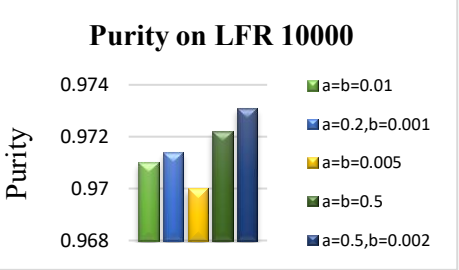
شکل ٤- بررسی معیار خلوص در پایگاه داده میزشام



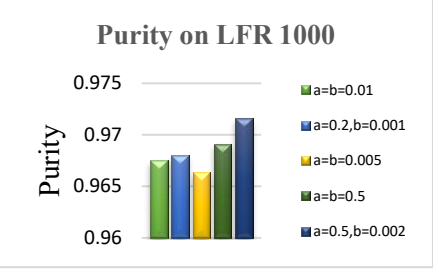
شکل۷- بررسی معیار خلوص در پایگاه داده LFR حاوی ۵۰۰ نود



شکل ٦- بررسی معیار خلوص در پایگاه داده LFR حاوی ۵۰ نود



شکل ۹- بررسی معیار خلوص در پایگاه داده LFR حاوی ۱۰۰۰۰ نود



شکل -۸- بررسی معیار خلوص در پایگاه داده LFR حاوی ۱۰۰۰ نود

۶-۱-۱- بررسی و تحلیل آزمایش

با توجه به اشکال و مقادیر حاصل شده، می‌توان گفت که خوشه‌های به‌دست آمده، تقریبا خالص هستند. به این معنی که هر خوشه شامل نودهای منحصر بفرد بوده و به ندرت شاهد نودهایی هستیم که نادرست خوشه‌بندی شده باشند و همچنین با افزایش میزان پاداش و کاهش میزان جریمه، نتایج بهتری حاصل می‌شود، بدین دلیل که افزایش پارامتر a منجر به دقت و سرعت بالاتر خواهد شد.

۶-۲- بررسی معیار NMI [1]

هدف : بررسی میزان شباهت خوشه‌های بدست آمده با خوشه‌های اصلی با توجه به مقادیر مختلف پارامترهای یادگیری.معیار NMI میزان شباهت خوشه‌های بدست آمده با خوشه‌های واقعی را نشان می‌دهد، که عددی بین صفر و یک می‌باشد. بدیهی است هر چه این عدد به یک نزدیک‌تر باشد به معنای آن است که خوشه‌بندی به خوشه‌بندی واقعی نزدیک‌تر است.

[1] *Normalized Mutual Information*

این معیار با توجه به فرمول زیر محاسبه میگردد:

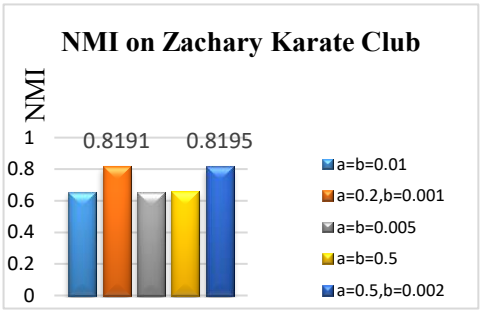
$$NMI(\Omega, \mathbb{C}) = \frac{I(\Omega,\mathbb{C})}{[H(\Omega)+H(\mathbb{C})]/2}$$

فرمول (۲-۶)

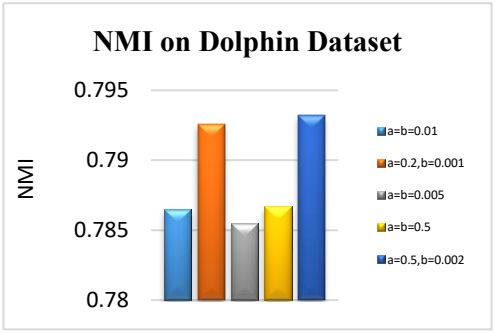
که در آن :

صورت کسر Mutual Information و مخرج کسر آنتروپی[1] می‌باشد (McDaid, Aaron., Derek Greene Neil Hurley,2011)

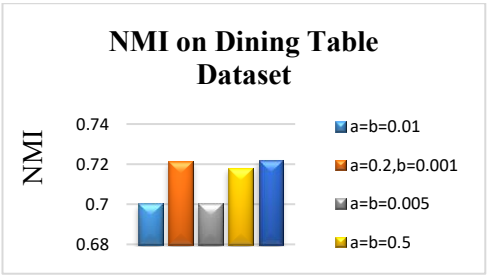
ما الگوریتم را برروی دادگان آزمایشی مندرج در جدول ۱ آزمایش کردیم و معیار NMI بدست آمده در هر پایگاه داده با توجه به پارامترهای تنظیم شده در ذیل به شرح شکل‌های زیر می‌باشد.



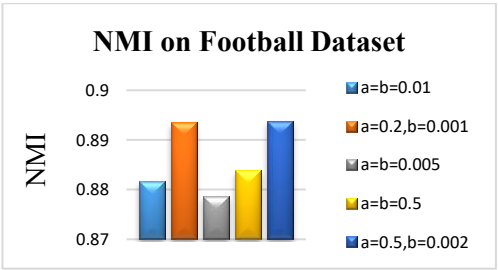
شکل ۱۰- بررسی معیار NMI در پایگاه‌داده باشگاه کاراته



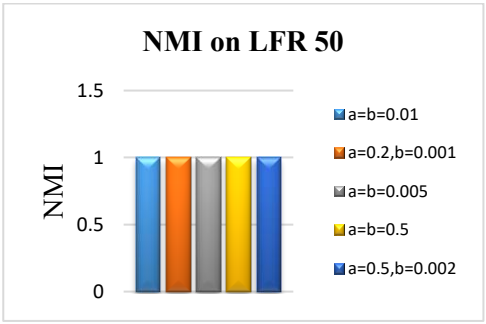
شکل ۱۱- بررسی معیار NMI در پایگاه‌داده دلفین‌ها



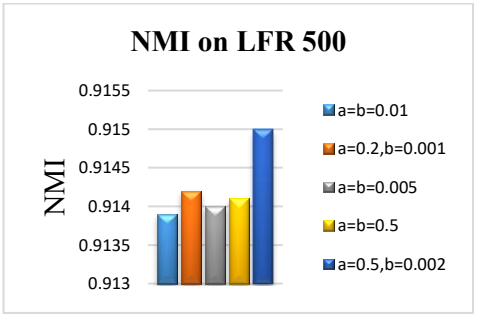
شکل۱۲- بررسی معیار NMI در پایگاه‌داده میزشام



شکل -۱۳- بررسی معیار NMI در پایگاه‌داده کالج فوتبال

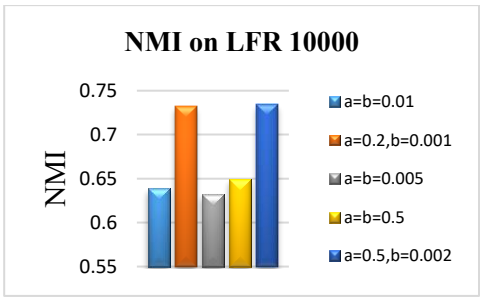


شکل -۱٤- بررسی معیار NMI در پایگاه‌داده LFR حاوی ۵۰ نود

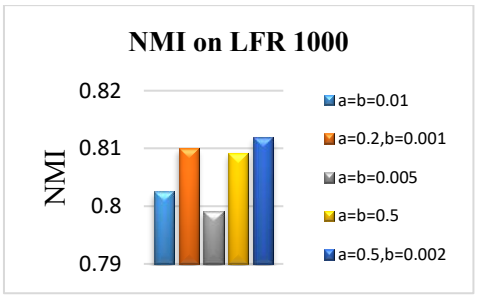


شکل -۱۵- بررسی معیار NMI در پایگاه‌داده LFR حاوی ۵۰۰ نود

[1] Entropy



شکل -۱۷- بررسی معیار NMI در پایگاه‌داده LFR حاوی ۱۰۰۰۰ نود



شکل -۱۶- بررسی معیار NMI در پایگاه‌داده LFR حاوی ۱۰۰۰ نود

با توجه به اشکال ۱۰ الی -۱۷ و مقادیر حاصل شده، می‌توان گفت که نودهای موجود در گراف‌ها، تقریبا درست خوشه‌بندی شده‌اند و افزایش در پارامتر پاداش منجر به افزایش دقت و یادگیری در اتوماتا شده است .به این معنی که هر خوشه تقریبا به خوشه اصلی شباهت دارد و خوشه‌بندی به خوشه‌بندی اصلی تقریبا نزدیک است. افزایش در پارامتر a منجر به نتایج دقیق‌تر و در مدت زمان کمتر خواهد شد.

۶-۳- بررسی معیار ماژولاریتی[1]

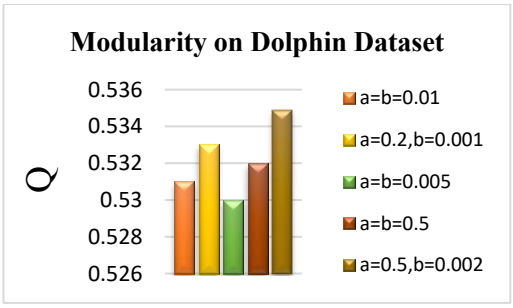
هدف : بررسی میزان مطلوبیت خوشه‌های بدست آمده

معیار ماژولاریتی نشان‌دهنده میزان مطلوبیت و کیفیت خوشه‌های بدست آمده است، که عددی بین صفر و یک می‌باشد. هر چه این عدد به یک نزدیک‌تر باشد یعنی خوشه‌های بدست آمده مطلوب‌تر هستند.

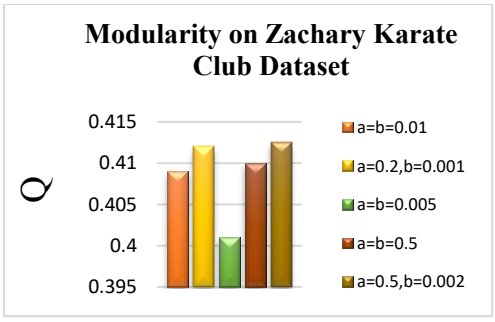
این معیار با توجه به فرمول زیر محاسبه می‌گردد.

$$Q = \sum_{i=1}^{c}(e_{ii} - a_i^2)$$ فرمول (۲-۷)

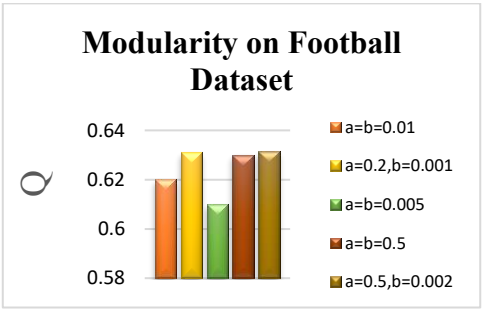
که در آن e_{ii} بخشی از رئوس در شبکه است که رئوس گروه i را به گروه j متصل می‌کندو $a_i=\sum_j e_{ij}$ می‌باشد. بنابراین $Q=\sum_i(e_{ii}-a_i^2)$ بخشی از رئوس است که در جوامع قرار گرفته است منهای ارزش مورد انتظار از همان مقدار.ما الگوریتم را برروی دادگان آزمایشی مندرج در جدول آزمایش کردیم و معیار ماژولاریتی بدست آمده در هر پایگاه داده با توجه به پارامترهای تنظیم شده در ذیل به شرح شکل‌های (۲-۱۸) الی (۲-۲۵) می‌باشد.

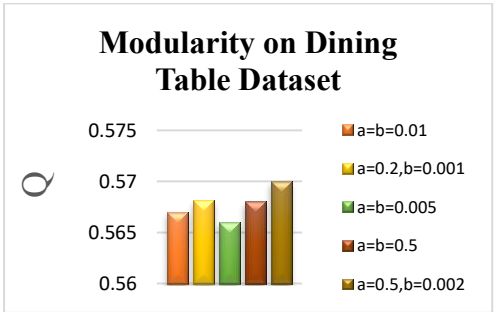


شکل -۱۹- بررسی معیار ماژولاریتی در پایگاه‌داده دلفین‌ها



شکل ۱۸- بررسی معیار ماژولاریتی در پایگاه‌داده باشگاه کاراته

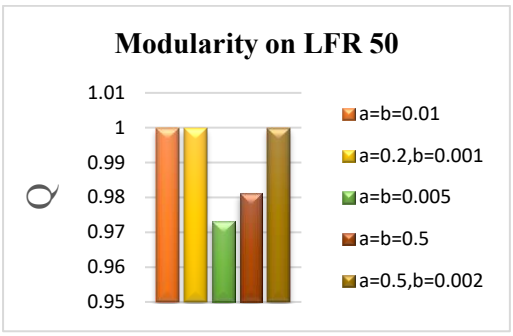




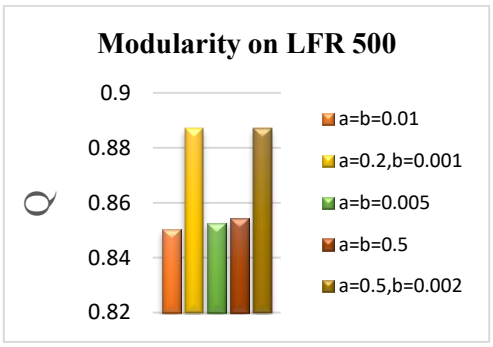
[1] Modularity

شکل -۲۰- بررسی معیار ماژولاریتی در پایگاه‌داده میزشام

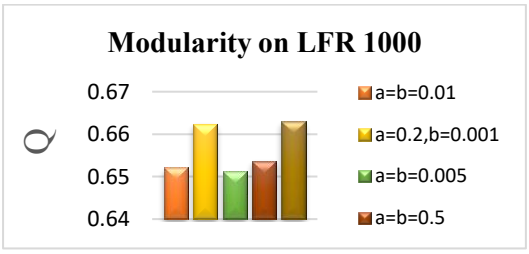
شکل -۲۱- بررسی معیار ماژولاریتی در پایگاه‌داده کالج فوتبال



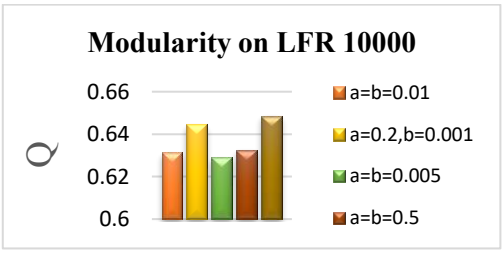
شکل -۲۲- بررسی معیار ماژولاریتی در پایگاه‌داده LFR حاوی ۵۰ نود



شکل -۲۳- بررسی معیار ماژولاریتی در پایگاه‌داده LFR حاوی ۵۰۰ نود



شکل -۲٤- بررسی معیار ماژولاریتی در پایگاه‌داده LFR حاوی ۱۰۰۰ نود



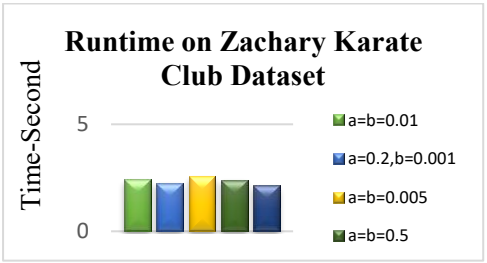
شکل -۲۵- بررسی معیار ماژولاریتی در پایگاه‌داده LFR حاوی ۱۰۰۰۰ نود

۶-۴- بررسی معیار زمان اجرا

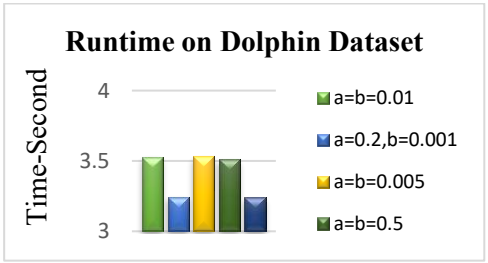
هدف : بررسی معیار زمان اجرای الگوریتم.

این معیار نشان‌دهنده مدت زمان لازم جهت اجرای کامل الگوریتم بر حسب ثانیه می‌باشد.

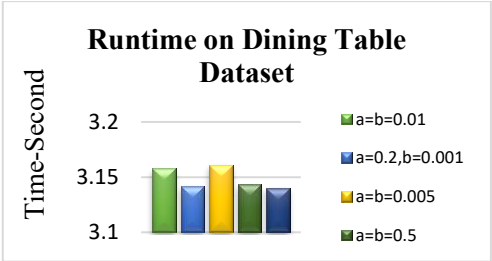
ما الگوریتم را برروی دادگان آزمایشی مندرج در جدول -۱ آزمایش کردیم و معیار زمان اجرای بدست آمده در هر پایگاه داده به شرح اشکال (۲۶) الی (۲۳۳) می‌باشد.



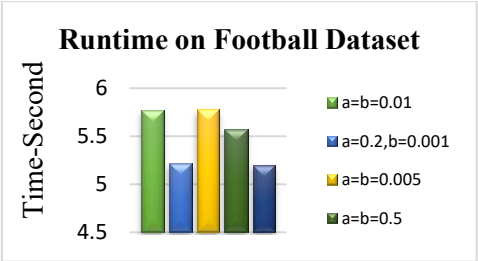
شکل -۲٦- بررسی معیار زمان اجرا در پایگاه‌داده باشگاه کاراته



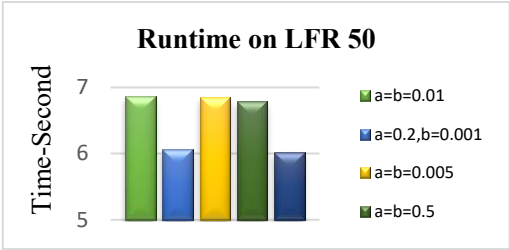
شکل -۲۷- بررسی معیار زمان اجرا در پایگاه‌داده دلفین‌ها



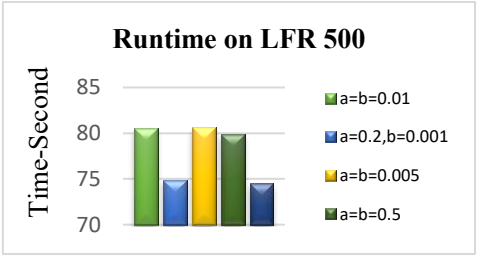
شکل ۲۸- بررسی معیار زمان اجرا در پایگاه‌داده میزشام



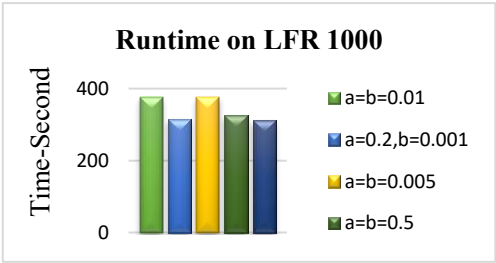
شکل ۲۹- بررسی معیار زمان اجرا در پایگاه‌داده کالج فوتبال



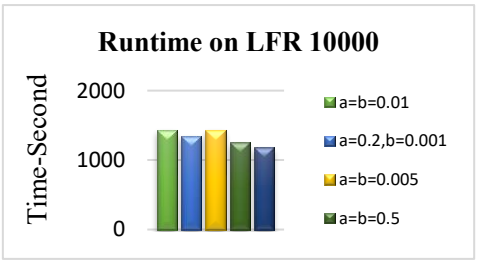
شکل ۳۰- بررسی معیار زمان اجرا در پایگاه‌داده LFR حاوی ۵۰ نود



شکل ۳۱- بررسی معیار زمان اجرا در پایگاه‌داده LFR حاوی ۵۰۰ نود

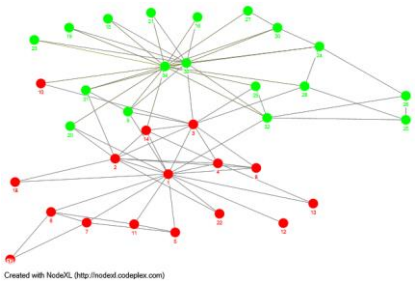


شکل -۳۲- بررسی معیار زمان اجرا در پایگاه‌داده LFR حاوی ۱۰۰۰ نود

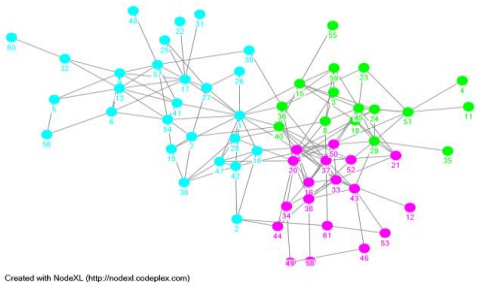


شکل -۳۳- بررسی معیار زمان اجرا در پایگاه‌داده LFR حاوی ۱۰۰۰۰ نود

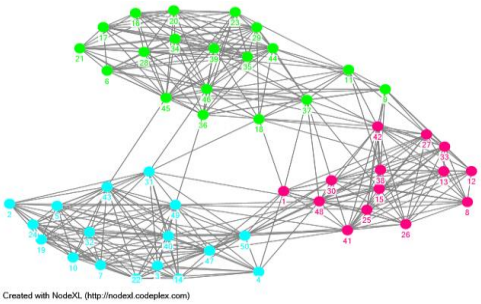
۷- خروجی‌های حاصل شده از اجرای الگوریتم



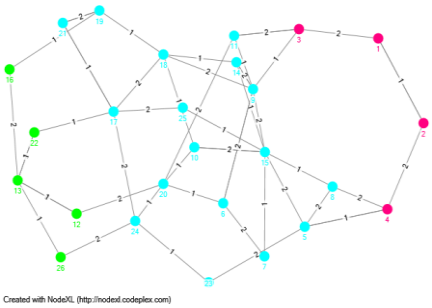
شکل -۳٤- اجتماعات یافت شده در پایگاه‌داده باشگاه کاراته



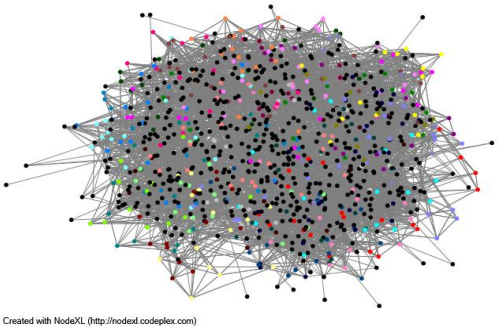
شکل -۳۵- اجتماعات یافت شده در پایگاه‌داده دلفین‌ها



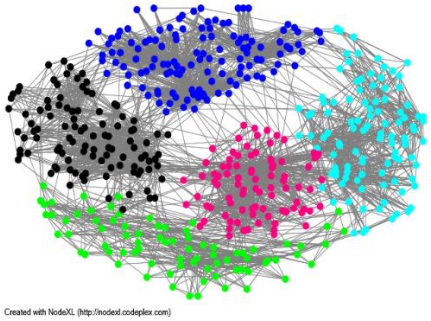
شکل ۳۷- اجتماعات یافت شده در پایگاه‌داده LFR حاوی ۵۰ نود



شکل ۳۶- اجتماعات یافت شده در پایگاه‌داده میز شام



شکل ۳۹- اجتماعات یافت شده در پایگاه‌داده LFR حاوی ۱۰۰۰ نود



شکل ۳۸- اجتماعات یافت شده در پایگاه‌داده LFR حاوی ۵۰۰ نود

جدول ۲-مقایسه مقدار ماژولاریتی الگوریتم‌های مختلف در سه مجموعه‌داده شبکه اجتماعی

نام الگوریتم	فوتبال	دلفین‌ها	باشگاه کاراته
GN	۰.۶۰۱۰	۰.۵۲۰۰	۰.۴۰۱۰
Blondel	۰.۶۰۴۶	۰.۵۲۷۷	۰.۴۱۹۸
LDF	۰.۶۰۲۸	۰.۵۱۷۹	۰.۴۱۹۸
OPT	۰.۶۰۴۶	۰.۵۲۸۵	۰.۴۱۹۸
CNM	۰.۵۷۷۰	–	۰.۳۸۰۰
WT	۰.۵۳۱۰	–	۰.۴۱۸۰
CFinder	۰.۵۹۶۰	–	۰.۱۸۳۰
MEP	۰.۶۰۱۰	–	۰.۳۷۱۰
QN	–	–	۰.۳۸۰۰
QEO	–	–	۰.۴۱۰۰
CDWDLA(a=b=0.01)	۰.۶۲۰۰	–	–
CDWDLA(a=0.2,b=0.001)	۰.۶۳۱۰	–	–
CDWDLA(a=0.005,b=0.005)	۰.۶۱۰۰	–	–
CDWDLA(a=0.5,b=0.002)	۰.۶۳۱۵	–	–
CDWDLA(a=b=0.5)	۰.۶۳۰۰	–	–
CDUDLA(a=b=0.01)	–	۰.۵۳۱۰	۰.۴۰۹۰
CDUDLA(a=0.2,b=0.001)	–	۰.۵۳۳۰	۰.۴۱۲۰
CDUDLA(a=0.005,b=0.005)	–	۰.۵۳۰۰	۰.۴۰۱۰
CDUDLA(a=0.5,b=0.002)	–	۰.۵۳۴۹	۰.۴۱۲۵
CDUDLA(a=b=0.5)	–	۰.۵۳۲۰	۰.۴۱۰۰

۸- نتیجه گیری

این پژوهش با هدف حل مسئله شناسایی اجتماعات، رویکردی نوین مبتنی بر اتوماتای یادگیر توزیع‌شده (مطابق با مبانی شیخ‌احمدی، دالوند، کشمیری، ۱۴۰۱) و ترکیب آن با روشی جدید مبتنی بر اشتراک همسایگی ارائه کرد. هدف، دستیابی به الگوریتمی با سرعت بالا و دقت بالا برای گراف‌های ایستا (وزن‌دار و بدون وزن) و گراف‌های بزرگ (با استفاده از روش‌های نمونه‌برداری ریحانیان، علیزاده، مینایی، ۱۳۹۲) بود.

الگوریتم‌های پیشنهادی یعنی CDUDLA (برای گراف‌های بدون وزن) و CDWDLA (برای گراف‌های وزن‌دار) پیاده‌سازی و ارزیابی شدند.

نتایج حاصل از معیارهای خلوص، NMI و ماژولاریتی نشان داد که این الگوریتم در اغلب موارد عملکردی رقابتی یا برتر نسبت به روش‌های مطرحی همچون GN (گیروان–نیومن)، بلوندل، LDF و OPT داشته است.

به‌عنوان نمونه، در دادگان کالج فوتبال، الگوریتم CDWDLA با ماژولاریتی ۶۳۱۵/۰، نتایج بهتری نسبت به تمام روش‌های مقایسه‌شده (بهترین روش دیگر با ماژولاریتی ۶۰۴۶/۰) نشان داد. همچنین در دادگان دلفین‌ها، این الگوریتم به ماژولاریتی ۵۳۴۹/۰ دست یافت که بالاتر از مقادیر گزارش‌شده برای الگوریتم‌های GN (۵۲۰/۰)، بلوندل (۵۲۷۷/۰) و LDF (۵۱۷۹/۰) بود.

تحلیل تأثیر پارامترهای پاداش (aaa) و جریمه (bbb) نشان داد که در حالت LRϵPL_{R\epsilon P}LRϵP (پاداش بیشتر از جریمه)، الگوریتم از دقت و سرعت بالاتری نسبت به حالت LRPL_{RP}LRP برخوردار است. این یافته نشان می‌دهد طراحی سیگنال تقویتی در اتوماتاهای یادگیر (در راستای دیدگاه Xiang, 2012) می‌تواند به‌عنوان اهرم کنترلی برای متوازن‌سازی دقت و سرعت در خوشه‌بندی عمل کند.

این پژوهش با چالش‌هایی از جمله محدودیت مقیاس (آزمایش تا ۱۰٬۰۰۰ نود)، عدم ارزیابی روی گراف‌های پویا (با وجود تأکید بر اهمیت پویایی شبکه‌ها توسط Tantipathananandh و Berger-Wolf, 2011)، وابستگی به تنظیم دستی پارامترها و عدم مقایسه با روش‌های یادگیری عمیق گرافی (مانند GNNها) مواجه بود. بر این اساس موارد زیر پیشنهاد شد:

۱. توسعه نسخه تطبیقی: تنظیم خودکار aaa و bbb (مثلاً با یادگیری تقویتی عمیق)

۲. ارتقای مقیاس‌پذیری: استفاده از پلتفرم‌هایی مانند Apache Spark یا نمونه‌برداری پیشرفته.

۳. پوشش گراف‌های پویا: توسعه الگوریتم افزایشی برای به‌روزرسانی بخش‌های متأثر از تغییرات گراف.

autorenewthumb_upthumb_down

منابع

- ریحانیان، علی و علیزاده، حسین و مینایی، بهروز،۱۳۹۲،تشخیص اجتماع مبتنی برموضوع گامی نوین درتحلیل شبکه های اجتماعی،همایش تخصصی بررسی ابعاد شبکه های اجتماعی،تهران،https://civilica.com/doc/225045
- شیخ احمدی، امیر و دالوند، شادی و کشمیری، احمد،۱۴۰۱،مروری بر شبکه های اجتماعی، کاربردهای آن و روشهای تشخیص انجمن،هفتمین کنفرانس بین المللی تحقیقات بین رشته ای در مهندسی برق، کامپیوتر، مکانیک و مکاترونیک در ایران و جهان اسلام،تهران،https://civilica.com/doc/1544106
- Delvenne, J. C., Yaliraki, S. N., & Barahona, M. (2010). Stability of graph communities across time scales. Proceedings of the National Academy of Sciences, 107(29), 12755-12760.
- Demetrescu, C., et al. (2010). Dynamic graph algorithms. In Algorithms and theory of computation handbook. Chapman & Hall/CRC.
- Dinh, T. N., & Thai, M. T. (2013). Community detection in scale-free networks: Approximation algorithms for maximizing modularity. IEEE Journal on Selected Areas in Communications, 31(6), 997-1006.
- Duan, D., Li, Y., Li, R., & Lu, Z. (2012). Incremental K-clique clustering in dynamic social networks. Artificial Intelligence Review, 38(2), 129-147.
- Fortunato, S. (2010). Community detection in graphs. Physics Reports, 486(3), 75-174.
- Harvey, N. J. A., Kleinberg, R., & Lehman, A. R. (2006). On the capacity of information networks. IEEE Transactions on Information Theory, 52(6), 2345-2364.
- Tantipathananandh, C., & Berger-Wolf, T. Y. (2011). Finding communities in dynamic social networks. In 2011 IEEE 11th International Conference on Data Mining (ICDM) (pp. 1236-1241). IEEE.
- Xia, Z., & Bu, Z. (2012). Community detection based on a semantic network. Knowledge-Based Systems, (26), 30-39.
- Xiang, J., & Hu, K. (2012). Limitation of multi-resolution methods in community detection. Physica A: Statistical Mechanics and its Applications, 391(20), 4995-5003.
- Yarahmadi, T., Akbari, J., & Zandevakili, F. (2012). A new method based on distributed learning automata for page ranking in web. International Journal of Physical Sciences, 7(13), 2066-2075.