نشریه علمی «پدافند الکترونیکی و سایبری»

سال هشتم، شماره ۱، بهار ۱۳۹۹؛ ص ۲۴-۱۷

روش ترکیبی تشخیص ناهنجاری با استفاده از تشخیص انجمن در گراف و انتخاب ویژگی

میثم میرزایی[1]، امین‌اله مه‌آبادی[2]*

۱- کارشناسی‌ارشد، دانشگاه امام حسین (ع)، تهران، ایران، ۲- استادیار، دانشگاه شاهد، تهران، ایران

(دریافت: ۹۷/۱۲/۱۴، پذیرش: ۹۸/۳/۲۸)

چکیده

تشخیص ناهنجاری یک موضوع مهم در بسیاری از حوزه‌های کاربردی شامل امنیت، سلامت و تشخیص نفوذ در شبکه‌های اجتماعی است. بیشتر روش‌های توسعه داده شده، فقط از اطلاعات ساختاری گراف ارتباطی یا اطلاعات محتوایی گره‌ها برای تشخیص ناهنجاری استفاده می‌کنند. ساختار یکپارچه بسیاری از شبکه‌ها از قبیل شبکه‌های اجتماعی این روش‌ها را با محدودیت مواجه ساخته است و باعث توسعه روش‌های ترکیبی شده است. در این مقاله، روش ترکیبی پیشنهادی تشخیص ناهنجاری مبتنی بر تشخیص انجمن در گراف و انتخاب ویژگی ارائه شده است که از ناهنجاری به‌عنوان اعضای ناسازگار در انجمن‌ها بهره برده و با استفاده از الگوریتم مبتنی بر تشخیص و ترکیب انجمن‌های مشابه، شناسایی گره‌های ناهنجار را انجام می‌دهد. نتایج آزمایش‌های تجربی روش پیشنهادی بر روی دو مجموعه از داده‌های دارای ناهنجاری واقعی، نشان‌دهنده قدرت تشخیص دقیق گره‌های ناهنجار و قابل مقایسه با آخرین روش‌های علمی است.

کلیدواژه‌ها: تشخیص ناهنجاری، شبکه‌های اجتماعی، داده کاوی، گراف‌کاوی

۱. مقدمه

تشخیص ناهنجاری شاخه‌ای از علم داده‌کاوی است که به دنبال یافتن نمونه‌های ناسازگار در مجموعه‌ای از داده‌ها است و همواره به عنوان یک موضوع مهم در تحلیل داده‌ها مطرح بوده و از قرن ۱۹ مورد مطالعه قرار گرفته است [۱]. تشخیص ناهنجاری می‌تواند باعث جلوگیری از افشای اطلاعات حساس، جلوگیری از دسترسی‌های غیرمجاز و پیش‌گیری از تصمیم‌گیری‌های خطا شود. تشخیص ناهنجاری در حوزه‌های فراوانی از قبیل تشخیص تقلب در کارت‌های اعتباری، بیمه، سامانه‌های سلامت، تشخیص نفوذ در شبکه‌های کامپیوتری و شناسایی هرزنامه در شبکه‌های اجتماعی کاربرد دارد. ناهنجاری‌ها نمونه داده‌هایی هستند که به میزان قابل توجهی با سایر نمونه داده‌ها متفاوت و ناسازگار هستند [۲]. ناهنجاری‌ها متناسب با حوزه کاربردی، پرت‌ها[1]، اختلالات، مشاهدات غیرواقعی و استثنائات نیز نامیده می‌شوند [۳]. منشا ناهنجاری‌ها می‌تواند رفتار کلاهبردارانه، خطای انسانی یا شکست سامانه‌ها باشد [۴].

همان‌گونه که بیان شد تشخیص ناهنجاری به‌دنبال یافتن نمونه داده‌هایی است که الگو یا رفتار معمول و متعارف موجود در مجموعه دسته‌بندی داده‌ها را نقض می‌کنند. البته مشخص کردن رفتار معمول به دلیل عدم وجود مرز مشخص و دقیق بین رفتار معمول و غیرمعمول، اصلی‌ترین چالش تشخیص ناهنجاری می‌باشد. تعریف‌های مختلف از رفتار معمول در حوزه‌های کاربردی مختلف، تلاش افراد خرابکار برای نزدیک کردن رفتار خود به رفتار معمول، تغییر تعریف رفتار معمول با گذشت زمان، وجود نویز در داده‌ها و در دسترس نبودن مجموعه‌های داده برچسب‌گذاری شده از دیگر چالش‌های این حوزه به شمار می‌آیند.

تعاملات اجتماعی، حمل‌ونقل، زنجیره تامین غذایی و بسیاری از پدیده‌ها قابل نمایش بصورت شبکه‌ای از عناصر به‌هم‌پیوسته هستند که از گراف برای نمایش آن‌ها استفاده می‌شود. برای مثال در گراف شبکه‌های اجتماعی، گره‌ها نشان‌دهنده افراد و یال‌ها نمایشگر تعاملات بین آن‌ها می‌باشند. بیشتر روش‌های توسعه‌داده‌شده تنها از اطلاعات محتوایی گره‌ها [۳] و یا اطلاعات ارتباطی بین آن‌ها [۵] برای تشخیص ناهنجاری استفاده می‌کنند. چالش اصلی دسته اول، سهولت دست‌کاری و جعل اطلاعات توسط افراد مخرب در بسیاری از حوزه‌های کاربردی از قبیل شبکه‌های اجتماعی است که باعث کاهش دقت این روش‌ها می‌شود. ساخت نمایه[2] بر اساس اطلاعات نادرست از سن، جنسیت، علاقه‌مندی و غیره درشبکه‌های اجتماعی از مصداق‌های این موضوع است.

هرچند اطلاعات مربوط به هم‌بندی و ارتباطات شبکه در

*رایانامه نویسنده پاسخگو: mahabadi@shahed.ac.ir

1- Outliers

2- Profile

اختیار کاربر نبوده و قابل دست‌کاری یا جعل نیست اما روش‌های مبتنی بر اطلاعات ارتباطی یا همان گراف‌های ساده نیز با یک مساله مهم یعنی از دست‌دادن سایر اطلاعات باارزش مرتبط با گره‌ها مواجه هستند. این دلایل باعث توجه به روش‌های تشخیص ناهنجاری در گراف‌های باویژگی[1] شده است که در آن‌ها اطلاعات ارتباطی و محتوایی گره‌ها در کنار یکدیگر قابل دسترس بوده و می‌توان از آن‌ها به‌صورت همزمان برای شناسایی ناهنجاری‌ها استفاده کرد. بیشتر روش‌های این دسته، از انجمن‌های موجود در گراف برای یافتن ناهنجاری‌های دارای انحراف قابل توجه از رفتار معمول استفاده می‌کنند. با توجه به تعداد زیاد ویژگی‌ها در بسیاری از شبکه‌ها، استفاده از تمام ویژگی‌ها برای تشخیص ناهنجاری باعث بروز مشکل مشقت ابعاد[2][۶] خواهد شد که ضمن دشوار ساختن استخراج الگوهای سودمند، کارایی روش‌ها را نیز کاهش خواهد داد.

در این مقاله روشی ترکیبی برای بهبود تشخیص ناهنجاری در گراف‌های باویژگی، مبتنی‌بر تشخیص انجمن و انتخاب ویژگی ارائه شده است که از ویژگی‌های گره‌ها و تعاملات بین آن‌ها برای تعیین ناهنجاری‌ها استفاده می‌کند. با در نظر گرفتن این فرض که ناهنجاری‌ها سعی بر قرارگیری در انجمن‌های مختلف برای افزایش ارتباطات خود دارند؛ روش ارائه‌شده برخلاف سایر الگوریتم‌های موجود، از ترکیب انجمن‌ها به‌جای خوشه‌بندی غیرهمپوشان گراف که در آن هر گره تنها به یک انجمن تعلق دارد استفاده می‌کند. ما ناهنجاری‌ها را به‌عنوان یک عضو از شبکه معرفی می‌کنیم که از لحاظ ساختاری و رفتاری با سایر اعضای انجمن خود متفاوت هستند. روش پیشنهادی، تشخیص ناهنجاری را با فرمول‌بندی در سطح گراف و با قابلیت کاربرد در انواع شبکه‌ها و مستقل از مورد مطالعه خاصی انجام می‌دهد.

در ادامه، پس از مرور کارهای مرتبط با تشخیص ناهنجاری در گراف‌های باویژگی در بخش۲، روش پیشنهادی در بخش ۳ تشریح و در بخش ۴ نتایج آن بر روی دو مجموعه داده واقعی نشان داده شده و با آخرین کارهای مرتبط مقایسه شده است. همچنین در بخش ۵ ضمن نتیجه‌گیری، آخرین چالش‌ها و موضوعات مهم برای تحقیقات آینده نیز ارایه می‌شود.

۲. کارهای مرتبط

با توجه به محدودیت‌های ذکر شده بر روش‌های تشخیص ناهنجاری که تنها بر اطلاعات محتوایی [۳] یا گراف ساده ارتباطات [۵] مبتنی هستند؛ توجه به ارائه روش‌های ترکیبی مورد توجه قرار گرفته است. در این روش‌ها، گراف تعاملات و ویژگی‌های گره‌ها در کنار هم برای شناسایی گره‌های ناهنجار مورد استفاده قرار می‌گیرند. خوشه‌بندی[3] اساس بیشتر روش‌های تشخیص ناهنجاری در گراف‌های باویژگی است. روش ارائه شده در [۷] با درنظر گرفتن انجمن‌ها به عنوان یک زمینه، به‌دنبال یافتن ناهنجاری‌های انجمنی[4] است که الگوی رفتاری متفاوتی با سایر نمونه داده‌ها در آن انجمن دارند. فرض اصلی آن است که داده‌های معمول انجمن‌ها را تشکیل می‌دهند و ناهنجاری‌ها به صورت تصادفی تولید شده و از توزیع یکنواخت پیروی می‌کنند. رادار[5] [۸] روشی مبتنی بر تحلیل ماتریس باقی‌مانده برای تشخیص ناهنجاری در گراف‌های باویژگی بدون جهت است. برای این منظور ابتدا ماتریس ویژگی‌ها برای دستیابی به نمونه‌های نماینده بازسازی و پس از بازسازی مجدد ماتریس ویژگی‌ها، ماتریس باقی‌مانده محاسبه می‌شود. ناهنجاری‌ها دارای الگوهای متفاوتی با نمونه‌های معمول در ماتریس باقی‌مانده هستند. در این روش فرض می‌شود دو گره دارای پیوند خواهند بود اگر ویژگی‌های آن‌ها شبیه به هم باشد و در واقع الگوهای مشابهی در ماتریس باقی‌مانده داشته باشند. روش پیشنهادی در [۹] برای تشخیص ناهنجاری در گراف‌های با ویژگی‌های عددی است. ابتدا گره‌ها و یال‌ها مستقل از گراف، برحسب مقادیر برچسب‌ها توسط روش نزدیک‌ترین همسایه ارزیابی می‌شوند. سپس گره‌ها و یال‌های معمول با یک مقدار ثابت و ناهنجاری‌ها با امتیاز ناهنجاری خود در گراف قرار داده می‌شوند. سپس زیرساختارهای پرتکرار گراف استخراج و پس از فشرده‌سازی، امتیاز ناهنجاری زیرساختارها محاسبه و ناهنجاری‌ها معرفی می‌شوند. نویسندگان در [۱۰] روشی برای تشخیص ناهنجاری‌های انجمنی در شبکه‌های اجتماعی ارائه کرده‌اند که ابتدا گره‌ها را بر اساس ویژگی‌های آنها خوشه‌بندی و در مرحله بعد از اطلاعات ساختاری برای تشخیص ناهنجاری استفاده می‌کند. فرض اصلی برای تشخیص ناهنجاری آن که افراد در یک خوشه تعاملات بیشتری با یکدیگر دارند. بنابراین، گره‌هایی در گراف که تعداد پیوندهای آن‌ها با خوشه‌های دیگر بیشتر از پیوندهای آن‌ها با خوشه خود باشد ناهنجارتر به شمار می‌آیند. روش [۱۱] با تقسیم گراف به خوشه‌های متراکم از لحاظ ساختاری و همگن از لحاظ ویژگی، پرت‌های انجمنی یا انحراف از زیر گراف پرتکرار [۱۲] را به‌عنوان ناهنجاری در شبکه وبلاگ‌نویسی‌های کوچک معرفی می‌کند. در این روش یک شبکه دوبخشی بین کاربران و پیام ها برای نمایش تعاملات همگن بین موجودیت‌های یکسان و تعاملات غیرهمگن

1- Attributed Graphs
2- Curse of Dimensionality

3- Clustering
4- Community Outliers
5- RADAR

بین موجودیت‌های متفاوت ساخته می‌شود؛ سپس یک الگوریتم خوشه‌بندی مبتنی بر سه بعدی‌سازی ماتریس غیرمنفی[1] برای تشخیص کاربران و پیام‌های ناهنجار استفاده می‌شود.

همگن بودن به معنای همبستگی ویژگی‌های گره‌ها به ساختار گراف است که نتیجه آن اتصال گره‌های شبیه به هم از لحاظ ویژگی در گراف تعاملات خواهد شد. وجود ویژگی‌های تصادفی، غیرمرتبط و نویز باعث نقض فرض همبستگی ویژگی‌ها و ساختار در بسیاری از گراف‌های با ویژگی می‌شود که در صورت استفاده از تمام ویژگی‌ها تشخیص ناهنجاری دقیقی صورت نخواهد گرفت. این موضوع باعث توسعه روش‌هایی شده است که از تمام ویژگی‌ها برای تشخیص استفاده نمی‌کنند. یک روش آماری برای انتخاب زیرمجموعه‌های مناسب از ویژگی‌ها [۱۳] ارائه شده است که ایده اصلی در آن، یافتن ویژگی‌هایی است که بیشترین شباهت را با ساختار گراف داشته و گره‌های گراف مقادیر مشابه در این زیرمجموعه‌ها دارند. یک گره ناهنجار است اگر تفاوت زیادی نسبت به همسایگان خود در گراف در مقادیر ویژگی‌های انتخاب شده داشته باشد. روش پیشنهاد شده در [۱۴]، مبتنی بر تحلیل فضای ویژگی است که در آن زیرگراف‌های متراکم از گره‌ها با ارتباطات درونی زیاد و شباهت درون خوشه‌ای بالا تشکیل داده می‌شوند. سپس زیرمجموعه‌هایی از ویژگی‌ها که بیشترین وابستگی را با ساختار خوشه دارند به دست می‌آیند که گره‌های خوشه بیشترین شباهت را به یکدیگر در آن ویژگی‌ها دارند. پس از تشکیل زیرگراف‌ها و زیرمجموعه‌ها امتیاز ناهنجاری برای هر گره محاسبه می‌شود. ایده اصلی آن است که اشیا معمول تلاش به ایجاد خوشه با بسیاری از اشیا شبیه به هم دارند. در [۱۵] روشی مشابه برای تشخیص ناهنجاری ارائه شده است؛ با این تفاوت که به جای انتخاب زیرمجموعه‌هایی از ویژگی‌ها که نسبت به کل ساختار همبستگی داشته باشند، زیرمجموعه‌ای از ویژگی‌های مناسب در همسایگی هر گره شناسایی می‌شوند. در این روش ابتدا فضای محلی هر گره بر اساس شباهت‌های ساختاری تشکیل می‌شود. سپس ویژگی‌هایی که واریانس آن‌ها در فضای محلی کمتر از کل مجموعه داده باشد مناسب تشخیص و انتخاب می‌شوند. گره‌ای ناهنجار است که دارای انحراف قابل توجهی در مقادیر ویژگی‌ها نسبت به سایر گره‌ها در یک فضای متراکم از گره‌های شبیه به هم باشد. روش گلنس[2] [۱۶] با استخراج ویژگی‌های مناسب هر انجمن، ناهنجاری‌های انجمنی را شناسایی می‌کند. ابتدا گره‌های گراف بر اساس ویژگی‌های ساختاری خوشه‌بندی و در انجمن‌های مختلف دسته‌بندی می‌شوند. پس از انتخاب ویژگی‌های مناسب هر انجمن، امتیاز ناهنجاری گره‌ها بر اساس انحراف مقادیر ویژگی‌های آنها از میانگین ویژگی‌های هر انجمن محاسبه و گره‌های با بیشترین امتیاز ناهنجاری معرفی می‌شوند.

۳. روش پیشنهادی

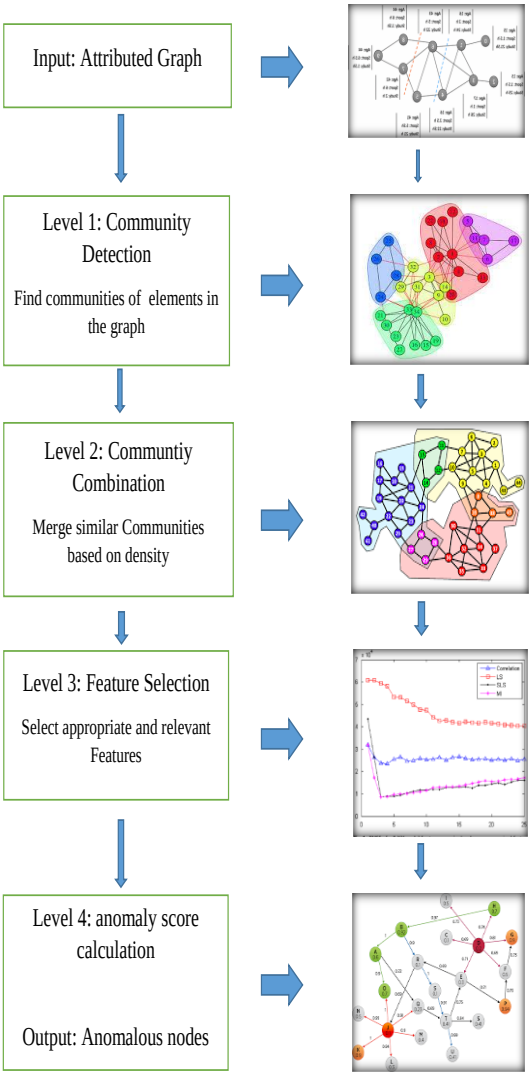
با بررسی و شناسایی محدودیت‌های روش‌های معرفی شده، روش پیشنهادی به دنبال تشخیص گره‌های ناهنجار با پیروی از رویکردهای مبتنی بر خوشه‌بندی، استفاده از ترکیب انجمن‌های مشابه و نیز انتخاب ویژگی است. این روش با توسعه الگوریتم گلنس [۱۶]، مطابق شکل (۱)، دارای چهار سطح تشخیص انجمن، ترکیب انجمن‌های شبیه به هم، انتخاب ویژگی و محاسبه امتیاز ناهنجاری گره‌ها است. ورودی روش یک گراف باویژگی و خروجی آن یک لیست مرتب شده از گره‌ها به ترتیب امتیاز ناهنجاری آن‌ها خواهد بود.

بالاتر بودن امتیاز هر گره به معنای ناهنجارتر بودن آن است. روش پیشنهاد شده دارای یک گام بیشتر نسبت به الگوریتم گلنس [۱۶] بوده که مربوط به ترکیب انجمن‌های مشابه است. هم‌چنین در الگوریتم گلنس [۱۶] امتیاز ناهنجاری گره‌ها تنها به ویژگی‌های محتوایی آن‌ها وابسته است و ویژگی‌های ساختاری گره‌ها در محاسبه‌دسته‌بندی این امتیاز نقشی ندارند؛ اما در روش پیشنهادی، امتیاز ناهنجاری ترکیبی از این دو دسته ویژگی است.

۳-۱. تشخیص انجمن

تشخیص انجمن مساله تقسیم گراف به خوشه‌هایی از گره‌های متصل به یکدیگر است. روش لووین [۱۷] یک الگوریتم ساده و کارآمد برای تشخیص انجمن در شبکه‌های بزرگ است. این روش یک ساختار انجمنی سلسله مراتبی کامل برای شبکه به نمایش می‌گذارد. محدودیت اندازه گراف ورودی در این روش تنها به محدودیت حافظه بستگی دارد و وابسته به محدودیت زمانی نیست؛ به همین دلیل این الگوریتم برای گراف شبکه‌های بزرگ همچون شبکه‌های اجتماعی مناسب است که می‌توانند شامل صدها میلیون گره باشند. ورودی الگوریتم لووین یک گراف (V,E) است. V و E مجموعه گره‌ها و یال‌های گراف هستند. تشخیص انجمن با تقسیم گراف به خوشه‌های $C=\{C1,C2,C3,\ldots Cn\}$ انجام می‌شود و هر Ci یک انجمن نامیده می‌شود. این روش غیرهم‌پوشان است و در آن $Ci\cap Cj=\phi$ برای همه i≠jها و هر گره تنها به یک انجمن تعلق خواهد داشت.

1- Non-Negative Matrix Tri-Factorization (NMTF)
2- Glance



شکل (۱): معماری روش پیشنهادی تشخیص گره‌های ناهنجار

۳-۲. ترکیب انجمن

از آنجا که به دنبال یافتن رفتار ناهنجار گره‌ها هستیم، باید مکان‌ها یا انجمن‌هایی را بیابیم که گره‌ها در آن‌ها ناهنجارتر هستند و از خود رفتاری غیر معمول‌تری به نمایش می‌گذارند. از سوی دیگر ماهیت بسیاری از شبکه‌ها همچون شبکه‌های اجتماعی به گونه‌ای است که استفاده از انجمن‌های غیرهم‌پوشان نمی‌تواند ساختار آن‌ها را به خوبی به نمایش بگذارد. در این شبکه‌ها معمولاً گره‌ها عضو چند انجمن هستند. برای پاسخ به این موضوع بعد از تشخیص انجمن‌ها، در این مرحله انجمن‌های مشابه را با یکدیگر ترکیب می‌کنیم. برای این‌کار ابتدا چگالی انجمن‌ها از رابطه (۱) محاسبه می‌شود که در آن |E|، تعداد یال‌های یک انجمن و |V| تعداد گره‌های موجود آن انجمن است.

$$ensity = \frac{2|E|}{|V|(|V|-1)} \qquad (۱)$$

پس از محاسبه چگالی، یال‌های تبادل‌شده بین هر دو انجمن را شمارش می‌کنیم. انجمن‌هایی مشابه و قابل ترکیب تشخیص داده خواهند شد که چگالی انجمن ترکیبی از چگالی هریک از دو انجمن اولیه به تنهایی بیشتر باشد.

۳-۳. انتخاب ویژگی

استفاده از کل ویژگی‌های گره برای تشخیص ناهنجاری به علت وجود ویژگی‌های تصادفی و نامرتبط باعث کاهش کارایی الگوریتم می‌شود. بنابراین، در روش پیشنهادی، بعد از تشخیص انجمن‌های ساختاری و ترکیب انجمن‌های مشابه، ویژگی‌های مناسب با استفاده از امتیاز لاپلاسین[1] [۱۸] انتخاب می‌شوند. این روش مناسب‌ترین ویژگی‌ها را که دارای واریانس بالا در عناصر نزدیک به هم هستند به‌صورت بدون نظارت و با ساخت گراف هر ویژگی انتخاب می‌کند.

۳-۴. امتیاز گره‌های ناهنجار

پس از انجام مراحل قبل، گره‌ها در انجمن‌های مختلف قرار گرفته و ویژگی‌های مناسب و مرتبط انتخاب می‌شوند. طبق روش پیشنهاد شده، هر گره می‌تواند در انجمن‌های مختلفی قرار داشته باشد. در این مرحله امتیاز ناهنجاری هر گره محاسبه و گره‌های با امتیاز بالاتر به عنوان ناهنجاری معرفی خواهند شد. طبق فرض اصلی روش، گره‌هایی ناهنجار هستند که تفاوت قابل‌ملاحظه‌ای با سایر گره‌ها در انجمن خود داشته باشند. به همین دلیل انحراف گره از سایر گره‌ها محاسبه می‌شود. استفاده از ویژگی‌های ساختاری یا محتوایی به‌صورت منفرد می‌تواند مانع از کشف ناهنجاری‌های پیچیده شود که سعی در شبیه‌کردن خود به داده‌های معمول دارند. برای غلبه بر این مشکل در روش پیشنهادی دو امتیاز انحراف ساختاری و انحراف ویژگی تعریف‌شده و برای هر گره محاسبه می‌شوند. امتیاز انحراف ساختاری گره i در یک انجمن با SCi نشان داده‌شده و از رابطه (۲) به‌دست می‌آید.

$$SC_i = \frac{|Deg_i - Deg_j<0|}{|C|} \qquad (۲)$$

1-Laplacian score

که در آن، $|C|$ تعداد اعضای انجمن و $|Deg_i - Deg_j < 0|$ تعداد گره‌های در انجمن با درجه بزرگتر از i را مشخص می‌کند. اعضای یک انجمن به هم شبیه و دارای تعامل با یکدیگر هستند؛ بنابراین، هرچه امتیاز انحراف ساختاری یک گره بیشتر باشد نشان‌دهنده تعامل کمتر گره با سایر اعضای انجمن و ناهنجارتر بودن آن است. امتیاز انحراف ویژگی گره i در یک انجمن با SFi نشان داده می‌شود. برای محاسبه این امتیاز ابتدا آن را برای هر ویژگی انتخاب شده به‌صورت جداگانه از رابطه (۳) محاسبه می‌کنیم.

$$SF_{i.l} = \frac{|a_l(v_i) - a_l(v_j) > A_l(C_i)|}{|C|} \quad (۳)$$

که در رابطه شماره (۳)، l ویژگی انتخاب شده، $a_l(v_i)$ مقدار ویژگی برای گره i، $A_l(C_i)$ میانگین اختلاف مقادیر همان ویژگی در انجمن، $|C|$ تعداد اعضای انجمن و $|a_l(v_i) - a_l(v_j) > A_l(C_i)|$ تعداد اعضای انجمن که اختلافی بزرگتر از میانگین اختلاف‌ها با گره i دارند را مشخص می‌کند. بزرگ‌بودن مقدار محاسبه شده نشان‌دهنده انحراف بیشتر گره نسبت به سایر اعضای انجمن خواهد بود.

پس از محاسبه امتیاز انحراف برای همه ویژگی‌های انتخاب شده، امتیاز بیشینه به‌عنوان امتیاز ناهنجاری گره i محاسبه و در نهایت امتیاز ناهنجاری نهایی گره با S_i برحسب امتیازهای انحراف ساختاری و ویژگی آن از رابطه (۴) به‌دست می‌آید:

$$S_i = \frac{SF_i + SC_i}{2} \quad (۴)$$

مراحل روش پیشنهادی در الگوریتم (۱) نشان داده شده است. ورودی الگوریتم یک گراف با ویژگی و خروجی آن لیست مرتب شده از امتیاز ناهنجاری گره‌ها است. بعد از تشخیص انجمن‌ها در خط ۲، انجمن‌های قابل ترکیب در خط‌های ۳ تا ۶ شناسایی و ترکیب می‌شوند. در خط ۷ ویژگی‌های مناسب انتخاب شده و در خط‌های ۸ تا ۱۶ امتیاز ناهنجاری هر گره در انجمن‌های مختلف براساس انحراف ساختاری و ویژگی بدست می‌آید. در پایان و در خط‌های ۱۷ تا ۲۰ امتیاز ناهنجاری بیشینه هر گره محاسبه و گره‌های دارای امتیاز بالاتر به عنوان ناهنجاری معرفی می‌شوند.

الگوریتم (۱): الگوریتم پیشنهادی امتیازدهی گره‌های ناهنجار

Input: G // Attributed Graph
Output: R // Anomaly ranking of the Vertices

1 $R \leftarrow \emptyset$
2: C← Community Detection(G)
3: **Foreach** $Ci \,\&\, Cj \in C$ **do**
4: if (density(Cij) > density Ci) & (density(Cij) > density Cj)
5: Cij← Ci + Cj
6: **end for**
7: A ←FeatureSelection(C)
8: **Foreach** $Ci \in C$ **do**
9: PCi ←mean difference values of attributes from A in Ci
10: **Foreach** $Vi \in Ci$ **do**
11: $SC_i = \frac{|Deg_i - Deg_j < 0|}{|C|}$
12: $SF_{i.l} = \frac{|a_l(v_i) - a_l(v_j) > A_l(C_i)|}{|C|}$
13: $SF_i = \max(SF_{i.l})$
14: $S_i = \frac{SF_i + SC_i}{2}$
15: **end for**
16: **end for**
17: **Foreach** Vi
18: $R_i = max(S_i)$
19: **end for**
20: Return R

۴. نتایج و بحث

یکی از چالش‌های اساسی تشخیص ناهنجاری در گراف‌های با ویژگی از قبیل شبکه‌های اجتماعی، عدم دسترسی به مجموعه داده‌های واقعی با ناهنجاری‌های برچسب گذاری شده برای ارزیابی و مقایسه روش‌های پیشنهادی مختلف است [۷]. از راه‌حل‌هایی که برای حل این مشکل وجود دارد تولید مجموعه داده‌های مصنوعی[1] و

1-Synthetic Dataset

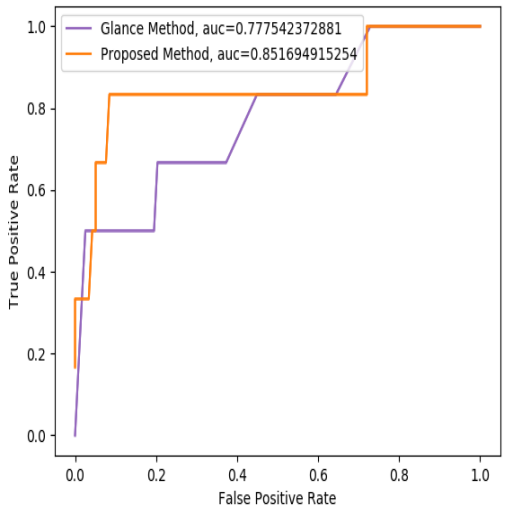
تزریق ناهنجاری به آن‌ها است. با توجه به هدف روش که تشخیص ناهنجاری روی گراف‌های با ویژگی است؛ استفاده از مجموعه داده‌های تصادفی مناسب نیست و تولید گراف و مجموعه ویژگی مرتبط کاری دشوار است. برای رفع این مشکل در روش [۱۴] مجموعه داده با ناهنجاری‌های برچسب خورده از شبکه فروش فیلم سایت آمازون وجود دارد. این مجموعه شامل ۱۲۴ گره و ۳۳۴ یال بین آن‌ها است. هم‌چنین هر گره دارای ۳۰ ویژگی عددی است. برچسب‌گذاری ناهنجاری‌ها بر اساس نظر نخبگان انجام شده است. برای جلوگیری از خطا در فرآیند برچسب‌گذاری ناهنجاری‌های واقعی، ابتدا خوشه‌بندی روی گره‌ها انجام شده و سپس از نخبگان خواسته شده است که در هر خوشه، ۱ یا ۲ عضو را به عنوان ناهنجار مشخص کنند. در نهایت به گره‌هایی که توسط حداقل ۵۰ درصد از نخبگان ناهنجار بوده‌اند، برچسب ناهنجاری زده شده است. مجموعه داده دوم نیز مربوط به شبکه فروش کتاب آمازون و شامل ۱۴۱۸ گره و ۳۶۹۵ یال با تعداد ۳۰ ویژگی است که با فرآیندی مشابه استخراج شده است.

برای ارزیابی کارآمدی، مقایسه بین روش پیشنهادی و سایر روش‌ها با استفاده از معیار AUC[1] انجام شده است. این معیار نشان‌دهنده سطح زیر نمودار ROC[2] است. هرچه مقدار این معیار برای یک دسته‌بند بزرگتر باشد کارایی نهایی دسته‌بند مطلوب‌تر ارزیابی می‌شود.

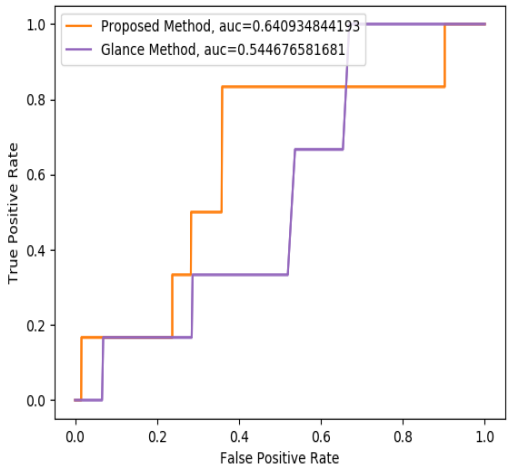
روش پیشنهادی با توسعه الگوریتم گلنس [۱۶] باعث بهبود نتایج شده است. این روش دارای یک گام بیشتر یعنی ترکیب انجمن‌های مشابه است. هم‌چنین علاوه بر ویژگی‌های محتوایی، در روش پیشنهادی یک امتیاز جدید مربوط به ساختار شبکه نیز تعریف شده است. شکل‌های (۲) و (۳) مقایسه نتایج این دو روش را از نظر مقدار AUC روی دو مجموعه داده واقعی نشان می‌دهد.

تعداد ویژگی‌های انتخابی در مجموعه داده اول ۲۰ درصد از ویژگی‌های مناسب و در مجموعه داده دوم ۳۰ درصد تعیین شده است. همان‌گونه که در شکل‌های (۲) و (۳) نشان داده شده است، روش پیشنهادی به ترتیب دارای مقدار ۸۵ درصد و ۶۴ درصد برای معیار AUC روی این دو مجموعه داده است در حالی‌که این مقادیر برای الگوریتم اصلی به ترتیب ۷۷ درصد و ۵۴ درصد هستند.

1-Area Under Curve
2-Receiver Operating Characteristic



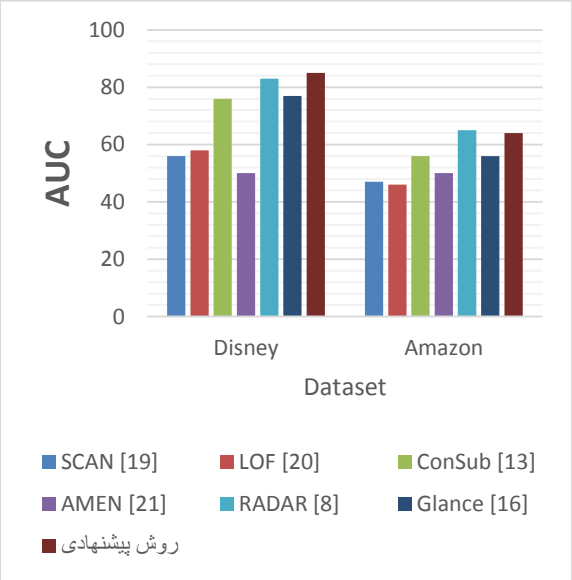
شکل (۲): مقایسه روش پیشنهادی و الگوریتم اصلی با مجموعه‌داده اول



شکل (۳): مقایسه روش پیشنهادی و الگوریتم اصلی با مجموعه‌داده دوم

همچنین برای مقایسه بهتر در شکل (۴)، مقایسه‌ای بین روش پیشنهادی و آخرین کارهای مرتبط با تشخیص ناهنجاری در گراف‌های باویژگی انجام شده است. برای ارزیابی بهتر، نتایج با سه دسته از روش‌ها مقایسه شده است. دسته اول روش‌های مبتنی بر گراف ساده هستند که تنها از ساختار شبکه برای تشخیص ناهنجاری‌ها استفاده می‌کنند[۲۰-۱۹]. دسته دوم روش‌هایی هستند که فقط براساس ویژگی‌های محتوایی ناهنجاری‌ها را شناسایی می‌کنند [۱۳ و ۲۱]. دسته سوم کارهای مشابه روش پیشنهادی هستند که از ویژگی‌های ساختاری و محتوایی به صورت همزمان برای این کار استفاده می‌کنند [۸] [۱۶]. البته روش‌های دسته دوم و سوم خود به دو حالت تقسیم می‌شوند: ۱. استفاده از تمام ویژگی‌ها یا ۲.

زیرمجموعه‌ای از ویژگی‌های مناسب. همانطور که نتایج نشان می‌دهد روش پیشنهادی از نظر معیار AUC در هر مجموعه داده دارای رتبه بالا در بین روش‌های توسعه داده شده است که نشان‌دهنده دقت بالای روش پیشنهادی در تشخیص ناهنجاری‌ها است.



شکل (۴): مقایسه روش پیشنهادی و سایر روش‌ها روی مجموعه داده‌های واقعی

از سوی دیگر با توجه به ماهیت روش پیشنهادی، امکان اجرای آن به صورت موازی وجود دارد. مراحل انتخاب ویژگی و تشخیص انجمن قابل انجام به صورت موازی هستند. پس از شناسایی انجمن‌ها مرحله محاسبه امتیاز ناهنجاری نیز می‌تواند بصورت موازی روی انجمن‌های مختلف انجام و در نهایت نتایج مشخص شود. بدین ترتیب امکان انجام اجرای موازی روش پیشنهادی می‌تواند باعث مقیاس‌پذیری آن در کار با شبکه‌های متوسط به بالا شود که تعداد انجمن در آن‌ها زیاد است.

۵. نتیجه‌گیری

در این مقاله از تحلیل گراف ایستای باویژگی برای تشخیص ناهنجاری استفاده شده است. هدف، یافتن گره‌های ناهنجار پیچیده‌ای است که سعی در شبیه‌سازی رفتار خود به نمونه‌های معمول در شبکه را دارند. برای شناسایی این ناهنجاری‌ها با استفاده از ترکیب انجمن‌های مشابه، سعی در یافتن مکان‌هایی است که می‌توانند ناهنجاری‌ها را به صورت بهتری مشخص کنند. هم‌چنین برای تعیین ناهنجار بودن یک گره انحراف آن نسبت به سایر هم انجمنی‌هایش از دو بعد ساختاری و محتوایی اندازه‌گیری شده است.

روش پیشنهادی توانسته است در تشخیص ناهنجاری‌ها بصورت موفق عمل کند و با توجه به نوع روش در سطح گراف شبکه، امکان استفاده از آن در حوزه‌های مختلف از جمله شبکه‌های اجتماعی، شبکه‌های رایانه‌ای و غیره که قابلیت نمایش به صورت گراف دارای ویژگی را دارند وجود دارد. چالش‌های نظری و عملی مهمی تشخیص ناهنجاری در گراف‌های باویژگی در کارهای آینده ما، تجمیع انتخاب ویژگی و تشخیص ناهنجاری در یک فرآیند، مقیاس‌پذیری روش برای گراف‌های بزرگ و ایجاد مجموعه داده‌های واقعی با اندازه‌های بزرگ مورد توجه هستند.

۶. مراجع

[1] F. Y. Edgeworth, "On discordant observations," Philosophical Magazine, Vol. 23, pp. 364-375, 1887.

[2] D. Toshniwal and S. Yadav, "Adaptive Outlier Detection in Streaming Time Series," in Proceedings of International Conference on Asia Agriculture and Animal, ICAAA, 2011.

[3] V. Chandola, A. Banerjee and V. Kumar, "Anomaly detection: A survey," ACM Computing Surveys (CSUR), Vol. 41, pp. 1-58, 2009.

[4] V. Hodge and Austin, "A survey of outlier detection methodologies," Intell, Vol, 22, p. 85–126, 2004.

[5] L. Akoglu, H. Tong and D. Koutra, "Graph based anomaly detection and description: a survey," Data Mining and Knowledge Discovery, Vol, 29, pp. 626-688, 2014.

[6] K. Beyer, J. Goldstien, R. Ramakrishnan and U. Shaft, "When Is "Nearest Neighbor" Meaningful?," International Conference on Database Theory, pp. 217-235, 1999.

[7] J. Gao, F. Liang, W. Fan, C. Wang and Y. Sun, "On community outliers and their efficient detection in information networks," in KDD, 2010.

[8] J. Li, H. Dani, X. Hu and H. Liu, "Radar: Residual Analysis for Anomaly Detection in Attributed Networks," in Internationl Joint Conference on Artificial Intelligence, pp. 2152-2158, 2017.

[9] M. Davis, W. Liu, P. Miller and G. Redpath, "Detecting Anomalies in Graphs with Numeric Labels," in Proceedings of the 20th ACM International Conference on Information and Knowledge Management, pp. 1197-1202, 2011.

[10] T. Ji, J. Gao and D. Yang, "A Scalable Algorithm for Detecting Community Outliers in Social Networks," in International Conference on Web-Age Information Management, pp. 434-445, 2012.

[11] W. Yang, G. W. Shen, W. Wang, L. Y. Gong and M. Yu, "Anomaly detection in microblogging via co-clustering.," Journal of Computer Science and Technology, pp. 1097–1108, 2015.

[12] C. C. Noble and D. J. Cook, "Graph-based anomaly detection," in Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data mining, pp. 631-636, 2003.

[13] P. I. Sánchez, E. Müller, F. Laforet and F. Keller, "Statistical Selection of Congruent Subspaces for Mining Attributed Graphs," in IEEE 13th International Conference on Data Mining, pp. 647–656, 2013.

[14] E. Müller, P. I. Sánchez , Y. Mülle and K. Böhm, "Ranking outlier nodes in subspaces of attributed graphs.," in IEEE 29th International Conference on Data Engineering Workshops (ICDEW),, 2013.

[15] P. I. Sánchez, E. Müller, O. Irmler and K. Böhm, "Local context selection for outlier ranking in graphs with multiple numeric node attributes.," in Proceedings of the 26th International Conference on Scientific and Statistical, 2014.

[16] M. A. Prado-Romero and A. Gago-Alonso, "Community Feature Selection for Anomaly Detection in Attributed Graphs," in Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications., pp. 109-116, 2017.

[17] V. D. Blondel, J. Guillaume, R. Lambiotte and Lef, "Fast unfolding of communities in large networks," 2008.

[18] X. He, D. Cai and P. Niyogi, "Laplacian Score for Feature Selection.," in Proceedings of the 18th International Conference on Neural Information Processing and Statistical, pp. 507-514, 2005.

[19] X. Xu, N. Yuruk, Z. Feng and T. A. Schweiger, "Scan: a structural clustering algorithm for networks.," in Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 824-833, 2007.

[20] M. Breunig, H. Kriegel, R. Ng, J. Sander and et al, "LOF: identifying density-based local outliers," in Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, pp. 93–104, 2000.

[21] B. Perozzi and L. Akoglu, "Scalable Anomaly Ranking of Attributed Neighborhoods," in SIAM International Conference on Data Mining, 2016.

Journal of Electronical & Cyber Defence
Vol. 8, No. 1, 2020, Serial No. 29

Hybrid Anomaly Detection Method Using Community Detection in Graph and Feature Selection

M. Mirzaei, A. Mehabadi*

Shahed University, Tehran

(Received: 05/03/2019, Accepted: 18/06/2019)

ABSTRACT

Anomaly detection is an important issue in a wide range of applications, such as security, health and intrusion detection in social networks. Most of the developed methods only use graph structural or content information to detect anomalies. Due to the integrated structure of many networks, such as social networks, applying these methods faces limitations and this has led to the development of hybrid methods. In this paper, a proposed hybrid method for anomaly detection is presented based on community detection in graph and feature selection which exploits anomalies as incompatible members in communities and uses an algorithm based on the detection and combination of similar communities. The experimental results of the proposed method on two datasets with real anomalies demonstrate its capability in the detection of anomalous nodes which is comparable to the latest scientific methods.

Keywords: Anomaly detection, Social networks, Data mining, Graph mining

* Corresponding Author Email: am@ipm.ir