استفاده از شبکه عصبی خودرمزگذار برای شناسای گره‌های ناهنجار در شبکه‌های اطلاعات

محمدحسین قنبری[۱]، مصطفی صالحی[۲]، وحید رنجبر[۳]

[۱] کارشناس ارشد رشته علوم تصمیم و مهندسی دانش، دانشکده علوم و فنون نوین، دانشگاه تهران ، تهران
mh.qanbari@ut.ac.ir

[۲] دانشیار، دانشکده علوم و فنون نوین، دانشگاه تهران ، تهران
mostafa_salehi@ut.ac.ir

[۳] استادیار، بخش فناوری اطلاعات دانشکده مهندسی کامپیوتر، دانشگاه یزد ، یزد
vranjbar@yazd.ac.ir

چکیده

برمبنای مفاهیم علوم داده و شبکه، هر سیستم اطلاعاتی در دنیای واقعی را می‌توان به شکل یک شبکه اطلاعاتی شامل عامل‌ها و ارتباط‌های مابین آنها مدلسازی کرد، که از آن جمله می‌توان به شبکه‌های اجتماعی، کامپیوتری، زیستی و اقتصادی اشاره کرد. تشخیص رفتارهای ناهنجار در شبکه‌های اطلاعاتی، کاربردهای مختلفی را متناسب با حوزه مورد مطالعه شامل می‌شود و از اهمیت ویژه‌ای برخوردار است. بر اساس تعریفی از ناهنجاری‌ها، که در این مقاله مورد توجه قرار گرفته است، منظور گره‌هایی هستند که رفتار متفاوتی نسبت به گره‌های همسایه‌ی خود دارند. روش‌های مختلفی برای شناسایی این رفتارها با استفاده از اطلاعات موجود در شبکه معرفی شده است، که در این بین روش‌های مبتنی بر تحلیل ساختاری، به دلیل بکارگیری اطلاعات پنهان در ساختار شبکه، توجه محققان را به خود جلب کرده‌اند. در این مقاله، با استفاده از ساختار شبکه و برمبنای رویکرد تحلیل طیفی، با بکارگیری سه فاز استخراج اطلاعات ساختاری، یادگیری شبکه عصبی به همراه کاهش بعد و رتبه‌بندی کاربران، به شناسایی ناهنجاری در شبکه‌های اطلاعاتی پرداخته شده است. با انجام آزمایش‌های مختلفی که روی مجموعه داده مصنوعی انجام شده، نتایج قابل قبولی در مقایسه با کارهای پیشین، از منظر شاخص ارزیابی نظارتی (شاخص مرتفع‌سازی)، بدست آمده است.

کلمات کلیدی

علم شبکه، تشخیص ناهنجاری ، شبکه‌های اطلاعات، شبکه‌های اجتماعی، تحلیل طیفی

۱- مقدمه

در سال‌های اخیر به دلیل گسترش شبکه‌های اجتماعی مانند فیسبوک و توئیتر، تحلیل شبکه‌های اجتماعی به یک حوزه پژوهشی روز دنیا تبدیل شده است که تشخیص ناهنجاری یکی از موضوعات این حوزه است. بطور کلی می‌توان ناهنجاری را رفتاری دانست که با رفتار سایر داده‌ها دارای تفاوت اساسی است [۱].

دو چالش اساسی در برخورد با شبکه‌های اطلاعاتی «استخراج اطلاعات در کنار ساختار» و «تحلیل شبکه با رویکرد ساختاری» هستند. در برخی از شبکه‌های اطلاعاتی، امکان استخراج اطلاعات به همراه ساختار، امری دشوار و هزینه‌بر است. به عنوان مثال، به دلیل حفظ حریم شخصی کاربران، امکان دسترسی به اطلاعات آن‌ها دشوار است. همچنین، در ساختار گرافی این شبکه‌ها، اطلاعات بسیاری نهفته است که با استفاده از آن‌ها می‌توان تحلیل‌های متفاوتی نسبت به اطلاعات کاربران بدست آورد.

امروزه با توجه به فراگیری شبکه‌های اجتماعی، حوزه‌های تحقیقاتی توجه چشمگیری به تشخیص ناهنجاری کرده‌اند. به دلیل وجود دیدگاه‌های ساختاری متفاوت در شبکه‌های اجتماعی، روش‌های گوناگونی برای تشخیص ناهنجاری روی این بسترها ارائه شده است [۲].

در این مقاله، با استفاده از ساختار شبکه و با بکارگیری سه فاز استخراج اطلاعات ساختاری، یادگیری شبکه عصبی به همراه کاهش بعد و رتبه‌بندی کاربران به بررسی و تحلیل ناهنجاری پرداخته شده است. در فاز اول، از روی ساختار گرافی شبکه، شباهت بین افراد با استفاده از معیار شباهت کسینوسی استخراج می‌شود. سپس، این داده‌ها به عنوان ورودی شبکه‌ی عصبی خودرمزگذار برای یادگیری و همچنین کاهش ابعاد ماتریس شباهت استفاده شده است. لایه مخفی میانی شبکه‌ی عصبی، به عنوان بردار ویژگی کاهش داده شده در نظر گرفته می‌شود. در فاز رتبه‌بندی، با کمک ساختار شبکه، فاصله‌ی بردارهای ویژگی تبدیل شده هر گره با بردارهای ویژگی همسایه‌هایش محاسبه می‌شود که به عنوان معیاری از ناهنجار بودن آن گره در نظر گرفته می‌شود. برای ارزیابی روش پیشنهادی نیز از یک روش نظارتی بر روی داده‌های مصنوعی استفاده شده است. برای این کار شبکه‌ی مصنوعی با پارامترهای مشخصی تولید می‌شود و سپس، به صورت تصادفی بخشی از شبکه ناهنجارسازی می‌شود.

در این مقاله، با توجه به چالش‌های «نبود اطلاعات کافی از گره‌ها و ارتباط‌ها» و «پیچیدگی ناهنجاری‌های ساختاری»، روش نوینی ارائه شده است که تنها با بکارگیری ساختار شبکه و با استفاده از شبکه خودرمزگذار، مدلی جهت تشخیص گره‌های ناهنجار ساخته می‌شود. ساختار شبکه به عنوان تنها اطلاعاتی که از شبکه‌های ایستای ساده در اختیار است، چالش عمده

تحلیل این دسته از شبکه‌هاست. در روش‌های تحلیل ناهنجاری در شبکه‌های ایستای ساده، با استفاده از استخراج ویژگی از ساختار شبکه و ارائه‌ی یک مدل از داده‌های ناهنجار، به تشخیص ناهنجاری‌ها پرداخته می‌شود.

در ادامه در بخش ۲ به مفاهیم پایه و معرفی شبکه‌های اطلاعاتی خواهیم پرداخت و سپس در بخش ۳ روش‌های موجود در این موضوع و چالش‌های تحلیلی روشها بررسی خواهند شد. پس از آن در بخش ۴ به معرفی راه حل پیشنهادی برای شناسایی گره‌های ناهنجار می‌پردازیم و در بخش ۵ روش پیشنهادی را با مورد ارزیابی قرار می‌دهیم. در نهایت در بخش ۶ نتیجه‌گیری و کارهای آتی را بیان می‌کنیم.

۲- مفاهیم اولیه

شبکه‌های اطلاعاتی با توجه به نوع ارتباط و موجودیت‌هایشان، آنها را می‌توان به دو دسته همگن و ناهمگن [۳] تقسیم کرد. در شبکه‌های ناهمگن، موجودیت‌ها و ارتباط بین آنها از یک نوع نیستند. به عنوان مثال، در شبکه‌ی ارجاعات مقالات، هر مقاله توسط یک ناشر چاپ می‌شود که نوع یال بین دو گره مقاله و ناشر، چاپ شدن است. همچنین، هر مقاله توسط یک یا چند نویسنده نوشته می‌شود که نوع یال بین دو گره مقاله و نویسنده، نوشته شدن است. مرجع [۴] انواع شبکه‌های ناهمگن به همراه تحلیل آنها را تشریح کرده است. برخلاف شبکه‌های ناهمگن، گره‌ها و یال‌ها در شبکه‌های همگن از یک نوع می‌باشند. به عنوان مثال، شبکه دوستی فیسبوک را در نظر بگیرید. گره‌ها نشانگر افراد و یال‌ها نشانگر رابطه دوستی بین آنهاست. در [۳]، شبکه‌های همگن به دو نوع شبکه‌های همگن پویا و ایستا تقسیم شده است. شبکه‌های پویا به شبکه‌هایی گفته می‌شود که گره‌ها و یال‌ها درحال تغییر هستند. مجموعه داده‌های شبکه‌های پویا شامل تصاویری از شبکه در زمان‌های مشخص است و از برچسب زمانی که در این مجموعه داده‌ها وجود دارد در تحلیل این نوع شبکه‌ها استفاده می‌شود.

شبکه‌های ایستا یکی از انواع شبکه‌های همگن هستند. در شبکه‌های ایستا، برخلاف شبکه‌های پویا تنها یک تصویر کلی از شبکه در یک زمان خاص در اختیار است. شبکه‌های ایستا را می‌توان بر اساس اطلاعات مازاد بر ساختار، به دو گروه شبکه‌های ایستای صفت‌دار و ساده تقسیم کرد [۳]:

- در شبکه‌های صفت‌دار، هر گره و یال شامل اطلاعاتی خارج از ساختار گرافی است. به عنوان مثال، در شبکه‌ی توئیتر برای هر فرد با توجه به نمایه‌ی او اطلاعاتی از قبیل سن، مکان جغرافیایی و یا علایق وی در اختیار است که برای تحلیل اطلاعات می‌توان از آنها استفاده کرد.
- در شبکه‌های ایستای ساده، جز ساختار گرافی شبکه اطلاعات دیگری از آن‌ها در دسترس نیست.

۳- کارهای پیشین

همانگونه که در بخش مفاهیم اولیه گفته شد شبکه‌های اطلاعاتی انواع مختلفی دارند و موضوع تشخیص ناهنجاری در هرکدام از این نوع شبکه‌ها مورد بررسی قرار گرفته است. به عنوان نمونه در [۵]، برای تشخیص ناهنجاری در شبکه‌های پویا، از روش LSTM که یک روش یادگیری عمیق است برای ساختن یک مدل شبکه‌ی عصبی استفاده شده است. در [۶] با محوریت صفات یال‌ها به تشخیص ناهنجاری پرداخته است. در این مقاله صفات مربوط به یال‌های شبکه را تحلیل کرده و گره‌های ناهنجار را شناسایی می‌کند.

کارهای بسیاری در زمینه تحلیل شبکه‌های اطلاعاتی با رویکرد مبتنی بر ساختار شبکه انجام شده است. روش‌های مبتنی بر تعبیه گراف [۷]، روش‌های مبتنی بر یادگیری ماشین [۸، ۹] و یادگیری عمیق [۱۰، ۱۱] از روش‌هایی هستند که در تحلیل شبکه‌های ایستای ساده نیز مورد استفاده قرار می‌گیرند.

روش‌های تشخیص ناهنجاری مبتنی بر ساختار در دو دسته تقسیم می‌شوند: روش‌های مبتنی بر ویژگی و روش‌های مبتنی بر نزدیکی. در روش‌های مبتنی بر ویژگی از روی ساختار گراف، ویژگی‌هایی جهت ساختن مدل تشخیص ناهنجاری، استخراج می‌شود. هدف این دسته از روش‌های ساختاری تبدیل فضای ساختاری گراف به فضای داده کاوی و تشخیص داده‌های پرت است. در روش‌های مبتنی بر نزدیکی از معیارهایی که از همسایگی و نزدیکی گره‌ها در ساختار گراف بهره می‌برند، استفاده می‌شود. در این مقاله، با استفاده از یک معیار شباهت برگرفته از نزدیکی گره‌ها در ساختار گراف، فضای ویژگی ساخته می‌شود. سپس این فضای ویژگی، با کمک ابزار شبکه‌ی عصبی خودرمزگذار کاهش بعد داده می‌شود.

در [۱۲]، با استفاده از مفهوم خود–شبکه (ego-net) و ویژگی‌های استخراجی از آن، روشی ارائه شده است که به تحلیل و بررسی ناهنجاری گره‌های شبکه می‌پردازد. خود–شبکه به زیرگرافی از گراف اصلی گفته می‌شود که شامل یک گره مرکزی و همسایه‌های آن است. یال‌های این زیرگراف متشکل از یال‌هایی از گراف اصلی است که هر دو رأس آن در مجموعه گره‌های خود–شبکه موجود باشد. ویژگی‌های استخراجی از خود–شبکه که در این روش معرفی شده است، شامل تعداد گره‌ها، تعداد یال‌ها، میزان بینابینی بودن گره مرکزی و واسطه بودن آن در خود–شبکه گره مرکزی است. سپس با تحلیل مقادیر این ویژگی‌ها الگوها و گره‌های ناهنجار امتیازدهی می‌شوند. برای این منظور، روش ارائه شده در [۱۲]، یک توزیع قانون توانی[۱] روی داده‌های استخراج شده برازش می‌کند. این برازش به این شکل انجام می‌شود، که یک نگاشت از تعداد یال‌ها به سایر ویژگی‌ها برای گره‌ها درنظر گرفته می‌شود. در نهایت سه تابع توزیع قانون توانی بدست می‌آید. مرحله بعد امتیازدهی گره‌هاست. امتیازدهی‌ها به این شکل انجام می‌شود که اختلاف مقادیر توزیع قانون توانی‌های برازش شده به ازای ورودی تعداد یال‌های خود–شبکه، با مقادیر ویژگی‌های تعداد همسایه‌ها، میزان بینابینی گره و میزان واسطه بودن آن، به عنوان امتیاز ناهنجاری آن گره محسوب می‌شود. علت درستی این روش از یک فرضیه در شبکه‌های واقعی حاصل می‌شود. در شبکه‌های واقعی انتظار می‌رود که ویژگی‌های خود–شبکه گره‌ها از توزیع قانون توانی پیروی می‌کنند. بنابراین گره‌هایی که ویژگی‌های مذکورشان با این توزیع‌ها اختلاف زیادی دارند، به عنوان ناهنجاری شناخته می‌شوند.

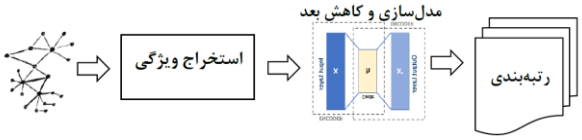
در مقاله [۱۳]، مقایسه‌ای بین «روش تحلیل مؤلفه‌های اساسی خطی و غیرخطی» و «شبکه عصبی خودرمزگذار یک لایه و عمیق» برای تشخیص ناهنجاری انجام شده است که نتایج نشان می‌دهند روش‌های غیرخطی عملکرد بهتری نسبت به روش‌های خطی دارند.

---

[۱] (Power-law)

۴- روش پیشنهادی

همانطورکه در شکل (۱) مشاهده می‌شود، روش پیشنهادی با استفاده از یک شبکه‌ی عصبی، ویژگی‌های استخراج شده از ساختار گراف را کاهش بعد داده و سپس با بکارگیری یک معیار امتیازدهی محلی، میزان ناهنجاری گره‌ها مشخص می‌شود. روش پیشنهادی از یک چارچوب سه مرحله‌ای بهره برده است که شامل مرحله‌ی استخراج ویژگی، کاهش بعد و معیار امتیازدهی محلی است.



شکل (۱) : ساختار شبکه عصبی خودرمزگذار

۴-۱- استخراج ویژگی

در این مرحله، از یک معیار شباهت به عنوان ویژگی برای هر گره استفاده می‌شود، که بیانگر این موضوع است که داده‌ها چه مقدار به یکدیگر شبیه هستند. شباهت موضوعی است که بسیار به فضا و کاربرد مسئله وابسته است.

در الگوریتم پیشنهادی از معیار شباهت کسینوسی برای تولید فضای ویژگی از روی ماتریس مجاورت استفاده شده است. از آنجایی که معیار فاصله اقلیدسی برای فضاهای پیوسته و چگال، معیار فاصله منهتن برای محاسبه اختلاف ویژگی‌ها و معیار شباهت جاکارد نیز برای محاسبه‌ی شباهت بین دو مجموعه مناسب هستند، نمی‌توان برای تولید فضای ویژگی از روی ماتریس مجاورت از آن‌ها استفاده کرد. اما معیار شباهت کسینوسی در فضاهایی با مقادیر مثبت کاربرد دارد تا مقدار شباهت در بازه [۰, ۱] محدود شود که در این مقاله بدلیل استفاده از ماتریس مجاورت بازه مقادیر شباهت در [۰, ۱] قرار می‌گیرد. یک دلیل دیگر برای انتخاب شباهت کسینوسی این است که این معیار در فضاهای برداری تنک بسیار کارامد است؛ و از آنجایی که ماتریس مجاورت ماتریسی تنک است، بهترین معیار برای استخراج ویژگی معیار شباهت کسینوسی است. دلیل تنک بودن ماتریس مجاورت ذات تنک شبکه‌های اجتماعی است.

۴-۲- مدل‌سازی و کاهش بعد

همانطور که گفته شد، بسیاری از روش‌های تشخیص ناهنجاری با ساختن یک مدل رفتاری از روی مجموعه داده، داده‌های ناهنجار را از سایر داده‌های مسئله متمایز می‌کنند. همچنین بزرگی شبکه‌های اجتماعی، باعث شده است که در روش پیشنهادی، از مدلی با دو هدف یادگیری الگوی رفتاری گره‌ها و کاهش بعد فضای ویژگی‌ها، استفاده شود. روش‌های کاهش بعد به دو دسته خطی و غیرخطی تقسیم می‌شوند. به عنوان مثال، تحلیل مؤلفه‌های اساسی یک روش برای کاهش بعد خطی و شبکه عصبی خودرمزگذار یک روش کاهش بعد غیرخطی است. در ادامه به این سؤال پاسخ داده می‌شود که «چرا شبکه عصبی خودرمزگذار بهتر از تحلیل مؤلفه‌های اساسی عمل می‌کند؟»

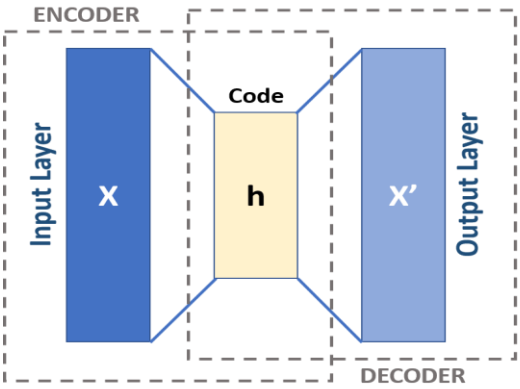
یکی از دلایل برتری روش‌های غیرخطی نسبت به روش‌های خطی انعطاف‌پذیری آنها برای برازش روی داده‌هاست. یکی از اشکالات روش تحلیل مؤلفه‌های اساسی این است که فضای ویژگی‌ها می‌بایست بصورت خطی قابل تبدیل باشند و تشخیص چنین موضوعی امری پیچیده و سخت است. اگر در شبکه عصبی خودرمزگذار از توابع فعال‌سازی غیرخطی استفاده شود، یک روش غیرخطی محسوب می‌شود که عموماً منظور از شبکه عصبی خودرمزگذار همین نوع غیرخطی آن است.

بطور کلی شبکه‌های عصبی در تبدیلات عملکرد بهتری دارند. علاوه بر این، با افزایش ابعاد داده‌ها روش تحلیل مؤلفه‌های اساسی پردازش کندتری نسبت به شبکه عصبی خودرمزگذار دارد. بنا بر نتایج تجربی موجود، هرچه فضای ویژگی داده‌ها غیرخطی باشد، خطای بازسازی روش‌های خطی مانند تحلیل مؤلفه‌های اساسی بیشتر از خطای بازسازی روش‌های غیرخطی مانند شبکه عصبی خودرمزگذار می‌شود.

در این مقاله، از یک شبکه عصبی خودرمزگذار غیرخطی جهت ساختن مدل و کاهش بعد استفاده شده است. علت استفاده از این شبکه عصبی در این مقاله این است که هم یک مدل کاهش بعد از فضای ویژگی داده‌ها و هم یک خوشه‌بندی مناسب برای تشخیص ناهنجاری ارائه می‌دهد.

شبکه عصبی خودرمزگذاری که در این مقاله مورد استفاده قرار گرفته شده است از یک تابع فعال‌سازی غیرخطی به نام تانژانت هایپربولیک بهره می‌برد. برتری تابع فعال‌سازی تانژانت هایپربولیک نسبت به تابع سیگموید این است که تابع تانژانت هایپربولیک مقادیر منفی نیز تولید می‌کند که باعث تمایز بهتری از داده‌ها در فضای حاصل می‌شود. همچنین از آنجایی که شبکه‌های اجتماعی دارای اجتماعات هستند، استفاده از مدل‌های غیرخطی کمک می‌کند تا خوشه‌بندی شبیه‌تری نسبت به اجتماعات موجود در شبکه حاصل شود. تجربیاتی که در جهت یافتن مدل بهتر در این مقاله کسب شده است، نشان می‌دهد که مدل، اجتماعات گره‌ها را در شکل خوشه‌هایی از داده‌ها در فضای کاهش بعد داده شده حفظ می‌کند.

ساختار شبکه عصبی خودرمزگذاری که در این پژوهش مورد استفاده قرار گرفته است، در شکل (۲) قابل مشاهده است که از خروجی بخش رمزگذار در فاز بعدی بهره برده شده است. در شکل (۲)، لایه ورودی (X) سطرهای ماتریس شباهت است. لایه مخفی (h) نیز فضای کاهش بعد داده شده را مشخص می‌کند. این شبکه عصبی سعی می‌کند تا در لایه خروجی (X') داده‌های لایه‌ی ورودی را تولید کند. پس از این که فرایند یادگیری مدل شبکه عصبی به اتمام رسید، از داده‌های تولید شده در لایه مخفی برای فاز رتبه‌بندی گره‌ها مورد استفاده قرار می‌گیرد. لایه‌های ورودی تا لایه مخفی میانی را بخش رمزگذار شبکه عصبی و لایه‌های مخفی میانی تا لایه خروجی را بخش رمزگشای شبکه عصبی می‌گویند.



شکل (۲) : ساختار شبکه عصبی خودرمزگذار

۴-۳- معیار امتیازدهی محلی

روش‌های تشخیص ناهنجاری به دو صورت می‌توانند نتایج را گزارش دهند: ارائه برچسب هنجار–ناهنجار و رتبه‌بندی ناهنجاری‌ها. در دسته اول، روش‌ها برای هر گره یک برچسب هنجار یا ناهنجار اعلام می‌کنند. اما در دسته دوم، روش‌ها گره‌های شبکه را به ترتیب از گره‌های ناهنجار به گره‌های هنجار گزارش می‌دهند، که در این مقاله نیز پس از امتیازدهی به گره‌ها یک رتبه‌بندی از آنها ارائه می‌شود.

خروجی‌های فاز مدل‌سازی و کاهش بعد شامل مجموعه داده‌ای است که به ازای هر گره، ویژگی‌های آن در فضای برداری تبدیل شده است. دو دستاوردی که از فاز قبل به دست آمده است عبارتند از: تبدیل شدن ماتریس شباهت با ابعاد بزرگ به ماتریس ویژگی‌ها با ابعاد کوچک و همچنین برجسته‌تر شدن داده‌های ناهنجار در حین مدل‌سازی و کاهش بعد. دستاورد اول این امکان را می‌دهد تا بتوان داده‌ها را مصورسازی کرد و یا هر تحلیل دیگری را با دقت خوبی روی فضای کوچک‌تر انجام داد. دستاورد دوم برجسته‌تر شدن داده‌های ناهنجار است. برای تفهیم این موضوع می‌بایست به دو نکته توجه کرد. نکته اول این است که در جریان یادگیری مدل سعی می‌شود تا خطای بازسازی به حداقل برسد و نکته دوم، نسبت بسیار زیاد داده‌های هنجار به داده‌های ناهنجار است که موجب می‌شود تا مدل، مقادیر ویژگی که برای داده‌های ناهنجار تولید می‌کند، از دقت بسیار کمتری نسبت به داده‌های هنجار برخوردار باشد. در واقع، به دلیل درصد کم داده‌های ناهنجار و عدم یکپارچگی در رفتار، شبکه عصبی نمی‌تواند به خوبی مدلی روی این داده‌ها برازش کند و در نتیجه، مدل مقادیر ویژگی‌های گره‌های ناهنجار را با دقت کمی تولید می‌کند.

سؤال اصلی این است که «چگونه متوجه این برجستگی داده‌های ناهنجار در بین سایر داده‌ها شویم؟». با پاسخ به این سؤال می‌توان فرمولی جهت تعیین داده‌های ناهنجار ارائه داد که ابتدا می‌بایست یکبار اطلاعات و ابزارهایی که در اختیار است مرور شوند. از شبکه، یک ساختار گرافی و از فاز قبل، مقادیر ویژگی‌های تبدیل شده هر گره در دسترس است. همانطور که گفته شد، از مدل شبکه عصبی خودرمزگذار خوشه‌بندی مناسبی از داده‌ها بدست می‌آید. می‌توان امتیاز ناهنجاری گره‌ها را میانگین فاصله اقلیدسی ویژگی‌های تبدیل شده هر گره تا همسایه‌هایش در نظر گرفت. دلیل کارایی این معیار این است که هر چه میانگین این فاصله‌ها به مقدار $2\sqrt{d}$ نزدیک‌تر باشد، داده در رتبه‌ی بالاتری قرار می‌گیرد. مقادیر امتیازها در بازه $[0, 2\sqrt{d}]$ است. زیرا بیشینه‌ی فاصله اقلیدسی دو بردار d بعدی (که دامنه مقادیر هر ویژگی در بازه $[+1,-1]$ قرار دارد) برابر است با $2\sqrt{d}$.

بدین ترتیب با تقسیم مقادیر میانگین فاصله‌ها به بیشینه فاصله‌ها، دامنه مقادیر در بازه $[0, 1]$ قرار می‌گیرند. روش پیشنهادی از ساختار شبکه‌های ایستای ساده ماتریس شباهت متقابل گره‌ها را استخراج می‌کند که رویکردی مبتنی بر نزدیکی گره‌ها است. علاوه بر این، یک روش طیفی‌ست که علاوه بر کاهش بعد، مدلی از داده‌های شبکه می‌سازد که به خوبی روی رفتار گره‌ها آموزش داده شده است. بدین ترتیب راهکار پیشنهادی، یک روش طیفی مبتنی بر ساختار است که با بکارگیری یک معیار نزدیکی به نام شباهت کسینوسی، مدلی جهت تشخیص ناهنجاری روی شبکه‌های همگن و ایستای ساده ارائه داده است. در بخش بعد مجموعه داده‌ی مورد استفاده، روش ارزیابی و نتایج حاصل ارائه می‌شود. همچنین این نتایج با یک روش تشخیص ناهنجاری نوین، تجزیه و تحلیل شده است.

۵- ارزیابی و نتایج

در ارزیابی روش، دو موضوع مورد توجه قرار گرفته است. موضوع اول مجموعه داده‌ای است که صحت عملکرد روش پیشنهادی روی آنها مورد بررسی قرار می‌گیرد. بطور کلی، مجموعه داده‌هایی که در ارزیابی‌ها مورد استفاده قرار می‌گیرند به دو دسته مجموعه داده‌های واقعی و مجموعه داده‌های مصنوعی تقسیم‌بندی می‌شوند. موضوع دوم معیار ارزیابی است. معیارهای ارزیابی را می‌توان به چهار دسته تقسیم کرد: ارزیابی نظارتی، ارزیابی بدون–نظارت، ارزیابی مقایسه‌ای و ارزیابی تفسیری. در ارزیابی‌های نظارتی از مجموعه داده‌هایی (واقعی یا مصنوعی) استفاده می‌شود که دارای برچسب باشند. در ارزیابی‌های بدون–نظارت، مجموعه داده‌ها فاقد برچسب هستند. به همین منظور، برای ارزیابی از تمرکز یا پراکندگی داده‌ها استفاده می‌کنند. در ارزیابی مقایسه‌ای پیش‌بینی‌های روش پیشنهادی با پیش‌بینی‌های روش‌های نوین مقایسه و نحوه عملکردشان تفسیر می‌شود. دسته آخر، ارزیابی‌های تفسیری هستند که نتایج پیش‌بینی شده را با توجه به اطلاعات موجود و یا مصورسازی داده‌ها مورد تفسیر و تحلیل قرار می‌دهند. به عنوان مثال، برای ارزیابی تفسیری تشخیص ناهنجاری در شبکه‌های اجتماعی می‌توان با مصورسازی بخش مرتبط شبکه، دلیل ناهنجاری گره‌های ناهنجار را مورد تحلیل و بررسی قرار داد.

۵-۱- مجموعه داده

در این مقاله، برای ارزیابی از مجموعه داده‌های مصنوعی استفاده شده است. برای ارزیابی با کمک مجموعه داده‌های مصنوعی، ابتدا با استفاده از روش‌های نوین، شبکه‌ای تولید می‌شود. سپس ناهنجاری‌هایی در شبکه تولید می‌شوند که از آن‌ها به عنوان برچسب ناهنجاری در ارزیابی روش استفاده می‌شود.

در این مقاله، از روش LFR[۲] [۱۴] جهت تولید شبکه‌های مصنوعی با پارامترهای متفاوت استفاده شده است. این روش از توزیع قانون توانی برای تولید درجات گره‌ها و اندازه اجتماعات استفاده می‌کند. بدین ترتیب می‌توان هر شبکه‌ای با توزیع درجه قانون توانی با پارامتر توان درجه دلخواه تولید کرد. روش LFR از پارامترهای دیگری نیز جهت تولید شبکه مصنوعی استفاده می‌کند: اندازه شبکه (N)، میانگین درجات (k)، بیشینه درجات (k_max)، پارامتر ادغام (μ)، کمینه تعداد اجتماعات (c_min)، بیشینه تعداد اجتماعات (c_max)، پارامتر توان درجه (λ) و پارامتر توان اندازه اجتماعات (β).

پارامتر μ، میزان لینک‌های بین اجتماعات را مشخص می‌کند. دامنه مقادیر این پارامتر در بازه $[0, 1]$ است. مقدار «۱» برای پارامتر ادغام بیانگر این است که صد درصد یال‌های گره‌ها بین اجتماعات قرار دارند. مقدار «۰» برای پارامتر ادغام یعنی تمام لینک‌ها درون اجتماعات قرار دارند. این پارامتر بیانگر میزان پیمانه‌ای بودن شبکه را نیز مشخص می‌کند. شبکه با میزان پیمانه‌ای بالا به این معناست که اجتماعات و یا خوشه‌ها به راحتی از یکدیگر قابل تمییز هستند. پارامتر λ، همان ضریب توان در توزیع درجه است. مقدار

[۲] Lancichinetti&Fortunato&Radicchi

این پارامتر در شبکه‌های اجتماعی معمولاً در بازه [۲, ۳] است. پارامتر β نیز ضریب توان در توزیع اندازه اجتماعات است که معمولا در شبکه‌های اجتماعی در بازه [۱, ۲] در نظر گرفته می‌شود.

پس از ساخته شدن شبکه تعدادی از گره‌ها به صورت تصادفی جهت ناهنجارسازی انتخاب می‌شوند. فرایند ناهنجارسازی بدین شکل است که همسایه‌های گره مورد بررسی با گره‌های تصادفی از شبکه جایگزین می‌شوند. در این صورت الگویی که در هنگام ساخت شبکه در نظر گرفته شده است، رعایت نخواهد شد. در روش LFR برای بررسی تأثیر چگال بودن شبکه روی روش کافیست میانگین درجات به بیشینه درجات نزدیک‌تر باشد.

۵-۲- معیار ارزیابی

همانطور که پیش‌تر گفته شد، چهار دسته معیار ارزیابی برای تحلیل و ارزیابی روش‌های پیشنهادی وجود دارد. در این مقاله با بکارگیری معیار نظارتی به ارزیابی روش پیشنهادی پرداخته شده است. برای این منظور، از روش شاخص مرتفع‌سازی برای ارزیابی شبکه‌های مصنوعی که توسط روش LFR تولید شده‌اند، مورد استفاده قرار گرفته است. این روش به دنبال جایگاه ناهنجاری در رتبه‌بندی حاصل از راهکار پیشنهادی است و بدین صورت که لیست رتبه‌بندی را به قطعه‌هایی تقسیم می‌کند و برای هر بخش امتیازی برای داده‌های ناهنجار درنظر می‌گیرد. هر چه ناهنجاری در قطعه‌های بالاتر قرار داشته باشد امتیاز بیشتری کسب می‌کند.

۵-۳- نتایج

در ادامه، نحوه عملکرد روش پیشنهادی و یک روش نوین تشخیص ناهنجاری ارائه شده است. مجموعه داده‌های مورد استفاده در ارزیابی نحوه عملکرد روش‌ها همان شبکه‌های تولید شده توسط روش LFR است. در شکل (۳)، تأثیر چگالی شبکه یا میانگین درجات در دقت عملکرد روش ارائه شده است. همانطور که قابل مشاهده است، روشی که در [۱۲] ارائه شده است عملکرد بهتری در شبکه‌هایی دارد که میانگین درجات از تنوع کمتری برخوردار است. با نزدیک شدن مقدار میانگین درجات به بیشینه درجات گره‌ها، شبکه هاب‌های خود را از دست می‌دهد و تنوع درجات در آن کاهش می‌یابد. کاهش تنوع درجات به این معنی است که گره‌های شبکه از درجات مشابهی تشکیل شده‌اند. اما در شبکه‌هایی که تنوع درجات در آنها بسیار بالا باشد، روش پیشنهادی عملکرد بهتری از خود نشان می‌دهد. در شبکه‌های اجتماعی واقعی، میانگین درجات گره‌های شبکه مقدار کوچکی نسبت به بیشینه آنها دارد.

همانطور که در شکل (۴) قابل مشاهده است، عملکرد روش‌ها با افزایش درصد ناهنجاری‌ها در شبکه کاهش می‌یابد. زیرا با افزایش درصد ناهنجاری‌ها و کاهش درصد گره‌های هنجار، رفتارهای هنجار و ناهنجار به سختی قابل تمییز هستند. نکته‌ی قابل توجه در شکل (۴) این است که میزان پراکندگی دقت‌های عملکرد (میله‌های خطا) برای روش پیشنهادی بیشتر از روش دیگر است. علت این امر وجود شرایط تصادفی در روش پیشنهادی در این مقاله است. اما دلیل وجود پراکندگی دقت برای روش ارائه شده در [۱۲]، شبکه‌های مصنوعی تولید شده است. تولید درجات گره‌ها، تولید اجتماعات شبکه و انتخاب تصادفی گره‌های کاندید جهت ناهنجارسازی، از دلایل وجود پراکندگی دقت روش‌ها است. همانطور که گفته شد، با افزایش درصد ناهنجاری دقت روش‌ها کاهش می‌یابد. در شکل (۴)، دقت روش‌ها با تغییرات درصد ناهنجاری‌های شبکه از ۱ درصد تا ۵۰ درصد ارائه شده است. اما در شبکه‌های واقعی درصد بسیار کمی از ناهنجاری‌ها وجود دارند که روش پیشنهادی در این مقاله، عملکرد بسیار خوبی را از خود نشان داده است.

در شکل (۵) مقایسه عملکرد روش پیشنهادی در این مقاله و روش تشخیص ناهنجاری ارائه شده در [۱۲] قابل مشاهده است. هر دو روش نسبت به خاصیت پیمانه‌ای ساختار شبکه حساس هستند. هر چه پارامتر ادغام در تولید شبکه مصنوعی کاهش یابد خاصیت پیمانه‌ای شبکه افزایش می‌یابد. با کاهش خاصیت پیمانه‌ای شبکه‌ها، تفکیک اجتماعات از یکدیگر به سختی انجام می‌شود. همچنین، با افزایش مقدار پارامتر ادغام، لینک‌های بین دو اجتماع افزایش می‌یابد و رفتارهای گره‌های اجتماع از یکپارچگی کمتری برخوردار خواهد شد. در این صورت مدل شبکه عصبی خودرمزگذار نمی‌تواند به خوبی روی رفتار گره‌های اجتماعات شبکه آموزش داده شود. در روش پیشنهادی، هر چه تعداد یال‌های بین دو اجتماع کمتر باشد، حذف یا تغییر یک یال بین این دو اجتماع کمتر به چشم می‌آید. در روش ارائه شده در [۱۲] نیز وقتی خاصیت پیمانه‌ای کاهش یابد نسبت تعداد گره‌ها، تعداد یال‌ها، مقدار ویژگی بینابینی و مقدار واسطه بودن در خود–شبکه گره مورد بررسی یکسان خواهد شد که در نهایت نمی‌تواند تمایزی بین رفتارهای هنجار و ناهنجار قائل شود. بنابراین خاصیت پیمانه‌ای شبکه‌ها رابطه مستقیمی با عملکرد هر دو روش دارد. بطور کلی در شبکه‌های اجتماعی، مقدار این پارامتر نسبتا کم است. همانطور که در شکل (۵) قابل مشاهده است، در مقادیر کوچک، دقت عملکرد روش پیشنهادی بسیار مناسب است.

همانطور که در نتایج قابل مشاهده بود، روش پیشنهادی این مقاله حساسیت زیادی به رفتارهای تصادفی دارد. هرگاه گره‌های شبکه، بدون درنظر گرفتن اجتماعی که به آن متعلق هستند، با سایر گره‌ها ارتباط داشته باشد، رفتار گره‌های شبکه تصادفی به نظر می‌آید. در شبکه‌هایی که گره‌های آن تعلق چندانی به اجتماعی نداشته باشند، تعریف اجتماع کمرنگ خواهد شد. پس همسایه‌های همه گره‌ها دارای اشتراک کمتری خواهند بود. این امر موجب می‌شود تا مدل نتواند الگوی مشخصی از رفتار گره‌های شبکه یاد بگیرد. این اتفاق زمانی‌که درصد ناهنجاری‌ها زیاد می‌شود نیز وجود دارد. در شبکه‌ای که درصد ناهنجاری‌ها زیاد باشد مدل شبکه عصبی خودرمزگذار نمی‌تواند الگوهای رفتاری شبکه را درک کند. زیرا با افزایش درصد ناهنجاری‌ها میزان تصادفی بودن رفتارهای گره‌های شبکه زیاد می‌شود. همچنین در خصوص میانگین درجات و پارامتر ادغام نیز این موضوعات صادق است. تأثیر میانگین درجات در تصادفی شدن رفتار گره‌های شبکه، زمانی است که با افزایش میانگین درجات، هاب‌ها از بین خواهند رفت و مدل نمی‌تواند الگوی رفتاری مشخصی را یاد بگیرد.

همچنین با افزایش پارامتر ادغام، شبکه خاصیت پیمانه‌ای خود را از دست می‌دهد و اجتماعات به یکدیگر شبیه می‌شوند. شبیه شدن دو اجتماع بدین معناست که یال‌های زیادی بین گره‌های آنها شکل خواهد گرفت. در این صورت گره‌ها جدا از تعلق به اجتماع خاص با سایر گره‌های شبکه ارتباط برقرار خواهند کرد. این موضوع باعث می‌شود تا مدل نتواند الگوی خاصی از شبکه یاد بگیرد. زیرا گره‌ها بصورت تقریبا تصادفی با یکدیگر در ارتباط خواهند بود و الگویی برای ارتباط آنها پیدا نخواهد شد.

[4] Shi C., Li Y., Zhang J., Sun Y., Yu P. S., *A Survey of Heterogeneous Information Network Analysis*, IEEE Trans. Knowl. Data Eng., 2017.

[5] Lin P., Ye K., Xu C. Z., *Dynamic network anomaly detection system by using deep learning techniques*, Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2019.

[6] Shah N., Beutel A., Hooi B., Akoglu L, Gunnemann S., Makhija D., Kumar M., Faloutsos C., *EdgeCentric: Anomaly Detection in Edge-Attributed Networks*, IEEE International Conference on Data Mining Workshops, ICDMW, 2017.

[7] Goyal P., Ferrara E., *Graph embedding techniques, applications, and performance: A survey*, Knowledge-Based Syst., 2018.

[8] Nickel M., Murphy K., Tresp V., Gabrilovich E., *A review of relational machine learning for knowledge graphs*, Proceedings of the IEEE. 2016.

[9] Eltanbouly S, Bashendy M, AlNaimi N, Chkirbene Z, Erbad A. *Machine Learning Techniques for Network Anomaly Detection: A Survey*. In2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT) 2020 Feb 2 (pp. 156-162). IEEE.

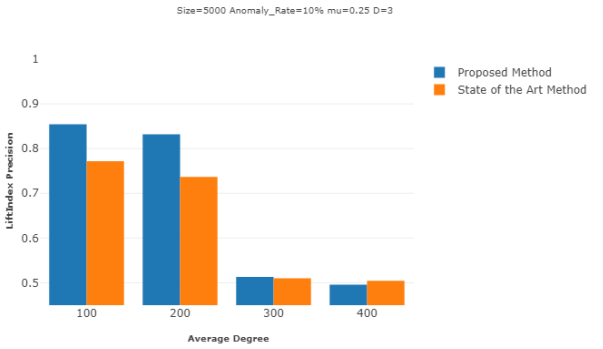
[10] Gamage S, Samarabandu J. *Deep learning methods in network intrusion detection: A survey and an objective comparison.* Journal of Network and Computer Applications. 2020 Nov, Vol. 169, No. 1, 2020.

[11] Bulusu S, Kailkhura B, Li B, Varshney PK, Song D. *Anomalous example detection in deep learning: A survey.* IEEE Access. Vol. 20, No. 8, 2020.

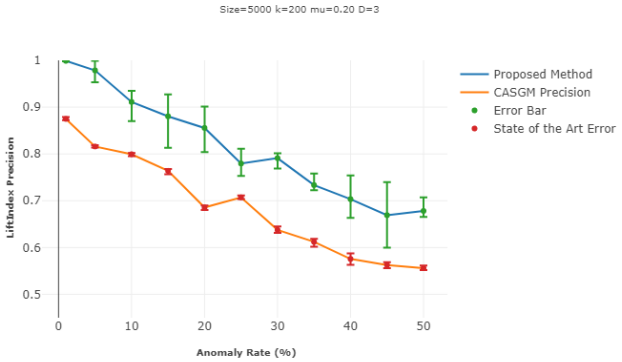
[12] Kaur R., Singh S., *A comparative analysis of structural graph metrics to identify anomalies in online social networks*, Comput. Electr. Eng., 2017.

[13] Sakurada M., Yairi T., *Anomaly detection using autoencoders with nonlinear dimensionality reduction*, ACM International Conference Proceeding Series, 2014.

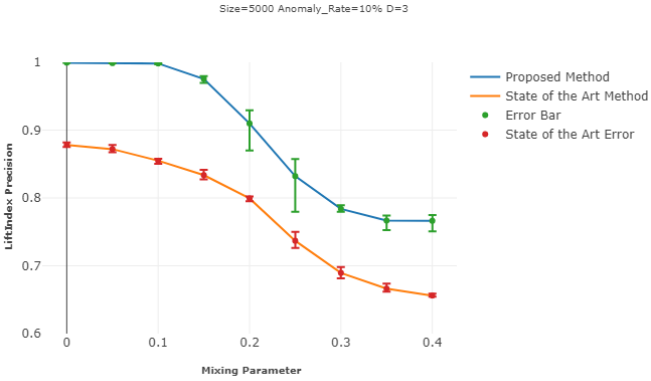
[14] A. Lancichinetti, S. Fortunato, and F. Radicchi, " Benchmark graphs for testing community detection algorithms," *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, 2008.



شکل (۳) : مقایسه روش پیشنهادی و روش نوین، متأثر از میانگین درجات (چگالی شبکه)



شکل (۴) : مقایسه عملکرد روش پیشنهادی و روش نوین، متأثر از درصد ناهنجاری



شکل (۵) : مقایسه عملکرد روش پیشنهادی و روش نوین، متأثر از پارامتر ادغام

مراجع

[1] Ganguli R, Mehta A, Sen S. *A Survey on Machine Learning Methodologies in Social Network Analysis.* In2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) 2020 Jun 4 (pp. 484-489). IEEE.

[2] Elghanuni RH, Ali MA, Swidan MB. *An Overview of Anomaly Detection for Online Social Network*. In2019 IEEE 10th Control and System Graduate Research Colloquium (ICSGRC) 2019 Aug 2 (pp. 172-177). IEEE.

[3] Akoglu L., Tong H., Koutra D., *Graph based anomaly detection and description: A survey*, Data Min. Knowl. Discov., vol. 29, no. 3, pp. 626–688, 2015.