
بسم الله الرحمن الرحيم

دانشگاه کردستان

دانشکده مهندسی

گروه مهندسی نرم افزار کامپیوتر

پایان نامه کارشناسی ارشد رشته مهندسی کامپیوتر گرایش هوش مصنوعی

عنوان:

روشی جدید برای تشخیص ناهنجاری یال بر اساس پیشگویی پیوند منفی

پژوهشگر:

سید آرمان حسینی

استاد راهنما:

دکتر فردین اخلاقیان

دکتر صادق سلیمانی

تابستان ۱۳۹۹

سپاسگزاری

سپاس **خدای** را که هر چه دارم از اوست.

به امید آنکه توفیق یابم جز خدمت به **خلق** او نکوشم.

تقدیم به **خانوادم**

مخصوصا

پدرم،کوهی استوار و حامی من در طول تمام زندگی

و

مادرم، سنگ صبوری که الفبای زندگی به من آموخت

تقدیم به **استارتاپ نوژن**

از اساتید گرامیم جناب آقای دکتر **فردین اخلاقیان** و دکتر **صادق سلیمانی** بسیار سپاسگذارم چرا که بدون راهنمایی‌های ایشان تامین این پایان نامه بسیار مشکل می‌نمود.

با تقدیم احترام

سید آرمان حسینی

۱۳۹۹/۰۷/۱۶

چکیده

تشخیص ناهنجاری در داده‌ها یک کار بسیار مهم و حیاتی است و کاربردهای زیادی در حوزه‌های مختلف از جمله امنیت، سلامت، امور مالی، مراقبت‌های بهداشتی و اجرای قانون دارد. در سال‌های اخیر روش‌های زیادی برای تشخیص ناهنجاری یا داده‌های پرت در مجموعه‌های بدون ساختار داده‌های چند بعدی ارائه شده است که بعضی از این روش‌ها روی ساختار گراف متمرکز شده‌اند. در این پایان نامه بر روی تشخیص ناهنجاری یال در گراف کار شده و دو روش بر اساس پیشگویی پیوند منفی برای تشخیص ناهنجاری یال پیشنهاد شده است. روش اول برای گراف‌های بدون وزن و روش دوم برای گراف‌های وزن‌دار ارائه شده و بر اساس عملکرد این روش‌ها، یال‌های ناهنجار در گراف با الگوریتم پیشگویی پیوند منفی تشخیص داده شده است. در دو روش پیشنهادی، از چهار الگوریتم پیشگویی پیوند، شاخص جاکارد، پیوست امتیازدهی، همسایه‌های مشترک و آدمیک-آدر به صورت بدون نظارت و مجزا استفاده شده است. همچنین از چهار مجموعه داده استاندارد دلفین، جاز، ایمیل و ترینیتی برای گراف‌های بدون وزن و از چهار مجموعه داده لسمیس[1]، پادشاه جیمز[2]، شبکه علمی[3] و نوجوان[4] برای گراف‌های وزن‌دار استفاده شده است. به منظور ارزیابی و کارایی روش پیشنهادی اول، چند درصد از کل یال‌های گراف، یال ناهنجار به گراف‌ها اضافه شد و با استفاده از روش پیشنهادی و هشت الگوریتم دیگر سعی شده که یال‌های ناهنجار تشخیص و نتایج روش‌ها باهم مقایسه گردد. نتیاج با معیاری‌های صحت، دقت، فراخوانی و معیار F1 ارزیابی شده است. برای روش پیشنهادی دوم، سعی شده است که با حذف یال‌های ناهنجار جوامع بهتری به وجود بیاید و جهت ارزیابی دو الگوریتم برچسب گذاری نامتقارن[5] و الگوریتم وزن‌دار بهینه‌سازی شده گروین-نیومن مورد استفاده قرار گرفته است. سپس برای تعیین بهبود جوامع از سه تابع کیفیت اجتماعات، ماژولاریتی، کارایی[6] و کاوریج[7]، استفاده می‌شود البته نیاز به ابداع روشی برای پیشگویی پیوند منفی در گراف‌های وزن‌دار و همچنین روشی برای اضافه کردن یال‌های ناهنجار به گراف‌های بدون وزن نیز وجود داشت که انجام شد.

کلمات کلیدی: تشخیص ناهنجاری، تشخیص ناهنجاری یال، تشخیص ناهنجاری در گراف، پیشگویی پیوند، پیشگویی پیوند منفی

---

[1] Lesmis

[2] King James

[3] Netscience

[4] Adolescent

[5] Asynchronous Label Propagation

[6] Performance

[7] Coverage

فهرست مطالب

عنوان **صفحه**

فهرست اشکال

فهرست جداول

اختصارات

Words of Group	Abbreviation
Link Prediction	LP
Negative Link Prediction	NLP
Posative Link Prediction	PLP
Common Neighbors	CN
Jaccard's Coefficient	JC
Adamic/Adar	AA
Preferential Attachment	PA
Onion Decomposition	OD
Modularity	Mod
Performance	Per
Coverage	Cov
Asynchronous Label Propagation Algorithm	AsynLPA
Label propagation algorithm	LPA
Random Walk Betweenness Centrality	RW_BC
Load Centrality	LC
Betweenness Centrality	BC
Katz	Ktz
Degree Centrality	DC
Eigenvector Centrality	EC
Closeness Centrality	CC
Betweenness Centrality	BC
Accuracy	ACC
Precision	PRE
Recall	REC
F-Score	F

فصل اول: مقدمه

۱-۱- موضوع پژوهش و تعریف مسئله

در دنیای امروز کشف ناهنجاری در داده‌ها کاربردهای زیادی در حوزه‌های مختلف از جمله امنیت، مراقبت‌های بهداشتی، امور مالی و . . . دارد. تشخیص ناهنجاری شاخه‌ای از داده کاوی است و در حوزه‌های مختلف تعریف‌های متفاوتی دارد. ناهنجاری تصادفی نیست یک انحراف اصلی از الگوی اصلی است. به عنوان مثال کسی که تقلب می‌کند سعی می‌کند رفتار خود را تا جایی که امکان دارد طبیعی نشان دهد. تشخیص ناهنجاری در گراف به این صورت تعریف می‌شود: اگر مجموعه‌ای از داده‌ها به صورت گراف نمایش داده شود، این قابلیت را دارد که فعالیت‌ها یا رفتارهای غیرقانونی را به وسیله تغییرات کوچک، مانند حذف و اضافه‌های که در گراف داده می‌شود، شناسایی کرد. در یک شناسایی مبتنی بر گراف چندین تغیر ممکن است رخ بدهد [۱]:

- یک گره غیر منتظره وجود دارد
- یک یال غیر منتظره وجود دارد
- یک برچسب متفاوت از انتظار روی گره وجود دارد
- یک برچسب متفاوت از انتظار روی یال وجود دارد
- یک گره مورد انتظار وجود ندارد
- یک یال مورد انتظار وجود ندارد

پیشگویی یال‌های تشکیل شونده در آینده، یا جعلی یا از قلم افتاده در شبکه‌هایی که با گراف نمایش داده می‌شوند، کاربردهای بسیار زیادی دارد. مساله پیشگویی پیوند در حالت پایه به صورت زیر تعریف می‌شود: اگر در زمان t یک تصویر لحظه‌ای از مجموعه لینک‌ها داشته باشیم، هدف پیشگویی پیوندها در زمان $t + 1$ است. معمولاً روش‌های پیشگویی پیوند به عنوان پیشگویی پیوند مثبت شناخته می‌شوند و یال‌ها یا ارتباطاتی را که در آینده به وجود می‌آیند را پیشگویی می‌کنند. در مقابل روش‌های کمی برای پیشگویی پیوند منفی وجود دارد، پیشگویی پیوند منفی یعنی پیشگویی یال‌ها یا ارتباطاتی که در آینده ناپدید می‌شوند.

مسئله و چالش اصلی این است که آیا پیشگویی پیوند در حوزه تشیخص ناهنجاری کاربرد دارد یا نه؟ حال ما در این پایان نامه با استفاده از الگوریتم‌های پیشگویی پیوند منفی سعی کردیم یال‌های ناهنجار را پیشگویی کنیم و به این سوال جواب بدهیم. برای اینکار دو روش را پیشنهاد داده‌ایم. روش اول برای گراف‌های بدون وزن و روش دوم برای گراف‌های وزن‌دار ارائه شده است.

۱-۲- اهمیت موضوع

کشف تخلف در ادارات و تراکنش‌های مالی، کشف گروه‌های تروریستی در شبکه‌های اجتماعی، کشف داده‌های پرت و یا پیش‌پردازش داده، کشف میکروب‌ها و جلوگیری از منتشر شدن یک بیماری یا به طور کلی تشخیص ناهنجاری در داده‌ها یکی از مهم‌ترین موضوعات در دنیای واقعی است. در این حوزه پژوهشگران روش‌های زیادی ارائه داده‌اند و این روش‌ها کاربردهای زیادی در دنیای امروزه دارد. برای همین ماهم به اهمیت این موضوع پی برده‌ایم و دو روش پیشنهادی برای تشخیص ناهنجاری در گراف ارائه داده‌ایم.

۱-۳- مفروضات پژوهش

۱) تشخیص ناهنجاری را بر یال در گراف اعمال می‌کنیم نه بر گره یا راس

۲) بنا به ضرورت الگوریتم‌های پیشگویی پیوند، قطر اصلی گراف باید صفر باشد

۳) معیار کارایی روش‌های کشف ناهنجاری، بهبود ساختار شبکه (بهبود قابلیت مؤلفه‌ای بودن) است

۱-۴- اهداف پژوهش

با توجه به چالش‌های پیش رو اهداف این پژوهش به شرح زیر است:

۱) ارائه روشی ساده برای تشخیص ناهنجاری یال در گراف از نظر یال‌های نابجا

۲) سنجش میزان کارایی پیشگویی پیوند در تشخیص ناهنجاری یال

نوآوری اصلی این پژوهش، اعمال رویکردی در پیشگویی پیوند برای اولین بار در حیطه‌ی تشخیص ناهنجاری یال می‌باشد که امکان پیاده‌سازی با الگوریتم‌های مختلف را دارد و ممکن است سبب شود دریچه‌های جدیدی از فعالیت‌های علمی در این زمینه گشوده شود. همچنین روشی جدید برای پیشگویی پیوند منفی در گراف‌های وزن‌دار و روشی جدید برای اضافه کردن یال‌های ناهنجار(دتیای نویر) جهت ارزیابی الگوریتم‌ها ارائه شده است.

۱-۵- روش تحقیق

اولین مرحله تهیه مجموعه داده‌های استانداردی از نوع گراف بود که باید تهیه می‌کردیم، برای انتخاب مجموعه داده یک محدودیت بزرگ داشتیم چون مجموعه داده‌های بزرگ سخت‌افزارهای قوی لازم داشتند. برای همین ما سعی کردیم که مجموعه داده‌های مناسبی با توجه به این محدودیت انتخاب کنیم. بعد الگوریتم‌های پیشگویی پیوند منفی و مثبت را پیاده سازی کردیم و این الگوریتم‌ها را روی مجموعه داده‌ها

اجرا کردیم و با توجه به اندازه گراف سعی کردیم درصد مناسبی از یال‌های ناهنجار را حذف کنیم، این درصد حذف شده برای مجموعه داده‌های مختلف متغییر است. در روش اول چند درصد یال ناهنجار(دیتای نویز) به گراف اضافه کردیم و با روش‌های پیشنهادی، این یال‌های ناهنجار را تشخیص داده‌ایم و با چهار معیار ارزیابی دقت روش پیشنهادی را سنجسده و نتایج را با هشت الگوریتم دیگر مقایسه کرده‌ایم. از روش پیشنهادی دوم جهت بهبود جوامع استفاده کرده و به صورت زیر ارزیابی کرده‌ایم: به ازای حذف درصدی از یال‌های ناهنجار ما با الگوریتم‌های تشخیص جوامع را روی گراف اعمال کردیم و خروجی این کار را با سه معیار ماژولاریتی، کارایی و کاوریج ارزیابی کردیم.

۱-۶- ساختار پایان نامه

متن این پایان‌نامه با احتساب فصل مقدمه در پنج فصل نگارش شده است .خلاصه‌ی مطالب اصلی هر فصل به شرح زیر است .در فصل ۲، ابتدا مفاهیم پایه‌ی تشخیص ناهنجاری، پیشگویی پیوند، توضحیات تکمیلی رساله و معیارهای ارزیابی بیان شده است. در فصل ۳، روش پیشنهادی اول و دوم برای مسئله تشخیص ناهنجاری یال در گراف‌های بدون وزن و وزن‌دار با جزئیات شرح داده شده است. در فصل ۴، در این فصل، مجموعه داده‌های استفاده شده برای گراف‌های وزن‌دار و بدون وزن ارائه شده است و نتایج دو روش پیشنهادی بر روی مجموعه‌های داده‌ها، در دو بخش مجزا ارائه و تفسیر خواهد شد. در فصل ۵، به جمع‌بندی کلی از روش‌های پیشنهادی در این پایان نامه پرداخته می‌شود.

فصل دوم: کلیات

۲-۱- مقدمه

این فصل شامل مباحثی است که از یک طرف پایه‌های اصلی این تحقیق محسوب می‌شوند و از طرف دیگر به درک بهتر موضوعات مطرح شده در ادامه رساله کمک می‌کنند. دو بخش اصلی این فصل مروری است بر ادبیات موضوعی در دو زمینه اصلی که عبارت‌اند از تشخیص ناهنجاری و پیشگویی پیوند. بخش اول درباره تشخیص ناهنجاری، کاربردها و انواع روش‌های آن است و بخش دوم به تشریح پیشگویی پیوند،کاربردها و روش‌های آن می‌پردازد. در نهایت، بخش سوم به توضیحات تکمیلی مورد نیاز برای ارایه روش‌های پیشنهادی اختصاص یافته است.

۲-۲- ناهنجاری[8]

تعاریف متعددی از ناهنجاری در مقالات متفاوت دیده می‌شود. این تعاریف با توجه به دامنه کاربردی و داده‌ها متفاوت می‌باشند. یکی از اولین تعاریف ارائه شده مربوط به هاوکینز در سال ۱۹۸۰ می‌باشد [۲]. هاوکینز ناهنجاری را به عنوان مشاهده‌ای که با اختلاف زیاد، متفاوت از مشاهدات دیگر است و با روش‌های متعدد ظن و بدگمانی نسبت به آن ایجاد می‌شود، تعریف می‌کند. همانطور که مشخص است این تعریف بسیار کلی می‌باشد و می‌تواند مفهوم ناهنجاری در زمینه‌ها و کاربردهای متفاوت را شامل گردد. حتی در زمینه‌های مختلف نام گذارهای متفاوتی به جای ناهنجاری استفاده می‌شود مانند داده‌های پرت[9]، مشاهدات ناسازگار[10]، استثنائات[11]، انحرافات[12] و شگفتی‌ها[13] [۲]. در مقاله دیگری ناهنجاری به این صورت تعریف شده است: ناهنجاری تصادفی نیست، یک انحراف کوچک از الگوی اصلی است. به عنوان مثال کسی که می‌خواهد تقلب کند سعی می‌کند که رفتار خود را تا جایی که امکان دارد طبیعی نشان دهد [۱].

در سال ۱۹۸۴ تعریفی توسط بارنت و لوئیس از ناهنجاری ارائه شد: "یک ناهنجاری به صورت قابل توجهی از نمونه‌های رخ داده دیگر، انحراف دارد". تعریف دیگری نیز توسط جانسون در سال ۲۰۰۲ به این مضمون ارائه شد: " ناهنجاری، مشاهده‌ای است در یک مجموعه

---

[8] Anomaly

[9] Outliers

[10] Discordant Observations

[11] Exceptions

[12] Aberrations

[13] Surprises

داده که به نظر می‌رسد با دیگرهای بخش‌های آن مجموعه داده ناسازگار باشد".

در زندگی روزمره ما ، روش‌های تشخیص ناهنجاری به طور صریح یا ضمنی برای تشخیص انحرافات از چیزی که نرمال یا مورد انتظار است استفاده می‌شود. به عنوان مثال همسایگان با دیدن رفتار غیرمعمول یک غریبه در اطراف یک منزل ممکن است دزدی را شناسایی کنند یا بانک‌ها با دیدن الگوهای نامعمول در پرداخت‌ها در حساب فرد، می‌توانند فعالیت‌های کلاهبردارانه را تشخیص دهند. مطالعه روش‌های تشخیص ناهنجاری به قرن بیستم بر می‌گردد که ابتدا توسط انجمن‌های آماری مورد بررسی قرار گرفت [3].

۲-۳- انواع ناهنجاری

ناهنجاری‌ها به طور کلی می‌توانند به سه دسته تقسیم شوند [۴]: ناهنجاری نقطه[۱۴]، ناهنجاری متنی[۱۵] و ناهنجاری جمعی[۱۶]

۲-۳-۱- ناهنجاری نقطه

یک نمونه داده تنها که به طور قابل ملاحظه‌ای نسبت به بقیه نمونه‌ها ناهنجار است را ناهنجاری نقطه‌ای می‌گویند. این ساده‌ترین نوع ناهنجاری است و تمرکز بیشتر تحقیقات روی این نوع ناهنجاری است. برای مثال در شکل ۲-۱ نقاط O_1، O_2 و همچنین نقطه O_3 خارج از منطقه نرمال قرار دارند بنابراین به این نقطه‌ها ناهنجاری می‌گویند چون با داده‌های عادی متفاوت می‌باشند.

به عنوان مثال از زندگی واقعی، تشخیص تقلب در کارت‌های اعتباری را در نظر بگیرید. فرض کنید که تمام معملات انجام شده با کارت اعتباری عادی است یعنی مجموعه داده‌ای که داریم با معاملات کارت اعتباری فرد مطابقت دارد. برای سادگی مثال فقط ویژگی خرج کردن با کارت اعتباری را در نظر بگیرید. معامله‌ای که مبلغ معامله شده در آن در مقایسه با دامنه عادی هزینه‌های شخص بسیار زیاد باشد را می‌توان به عنوان ناهنجاری در نظر گرفت.

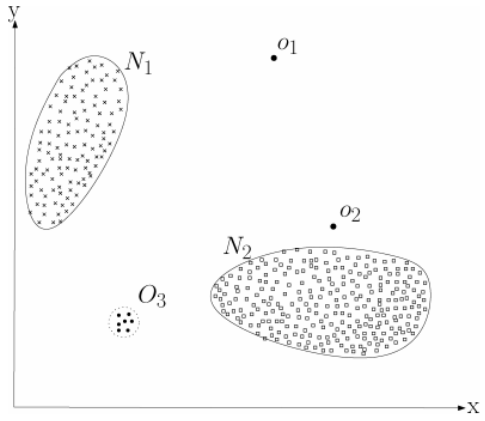
دراین تحقیق نیز تمرکز روی این نوع ناهنجاری می‌باشد. به طور کلی روش‌های تشخیص این نوع ناهنجاری می‌تواند به صورت بانظارت، بدون نظارت و نیمه نظارتی دسته‌بندی شود [۵]. [۶].

---

[14] Point Anomalies

[15] Contextual Anomalies

[16] Collective Anomalies



شکل ۲—۱ ناهنجاری نقطه در فضای دو بعدی، نقاط O_1، O_2 و O_3 ناهنجاری هستند

۲-۳-۲- ناهنجاری متنی

نوع دیگر ناهنجاری، ناهنجاری متنی می‌باشد که ناهنجاری شرطی نیز اتلاق می‌شود. تعریف این نوع ناهنجاری به این صورت است که، یک نمونه داده می‌تواند در یک زمینه خاص ناهنجاری باشد ولی در زمینه‌های دیگر ناهنجاری نباشد. برای مثال، در ژوئن پایین آمدن دما در بسیاری از کشورها طبیعی نیست اما همین قضیه در بعضی کشورها طبیعی است.

در ناهنجاری متنی مفهوم‌های ساختاری مجموعه داده باید به عنوان بخشی از فرمول مسئله مشخص شود. هر نمونه داده با استفاده از دو مجموعه ویژگی زیر تعریف می‌شوند:

- **ویژگی متنی**[۱۷]: از صفات متنی برای تعیین نوع نمونه متنی استفاده می‌شود. برای مثال در مجموعه داده‌های مکانی، طول و عرض جغرافیایی یک مکان از ویژگی‌های متنی است. در داده‌های سری زمانی، زمان یک ویژگی متنی است که موقعیت یک نمونه را در کل دنباله تعیین می‌کند.
- **ویژگی رفتاری**[۱۸]: ویژگی‌های رفتاری، ویژگی‌های غیرمتعارف یک نمونه را تعریف می‌کنند. به عنوان مثال، در یک مجموعه داده مکانی که میانگین بارندگی کل جهان را توصیف می‌کند، میزان بارندگی در هر مکان یک ویژگی رفتاری است.

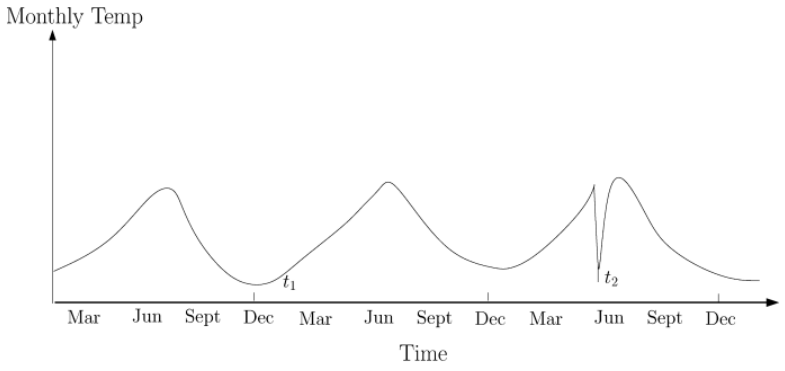
رفتار غیر عادی در این نوع ناهنجاری با استفاده از مقدار ویژگی‌های رفتاری در زمینه خاص تعیین می‌شود. یک نمونه داده ممکن

---

[17] Contextual Attributes

[18] Behavioral Attributes

است در یک زمینه ناهنجاری متنی باشد اما یک نمونه داده یکسان با همان ویژگی رفتاری در یک زمینه دیگر عادی تلقی شود. ویژگی‌های رفتاری و متنی در روش‌های تشخیص این نوع ناهنجاری خیلی مهم هستند. ناهنجاری متنی بیشتر در داده‌های سری زمانی و مکانی مورد بررسی قرار گرفته است. شکل ۲-۲ نمونه‌ای از یک سری زمانی مربوط به دمای ماهانه یک منطقه در چند سال گذشته را نشان می‌دهد دمای ۳۵ درجه فارنهایت ممکن است در طول زمستان (در زمان t_1) طبیعی باشد، اما همین مقدار در طول تابستان (در زمان t_2) یک ناهنجاری باشد.



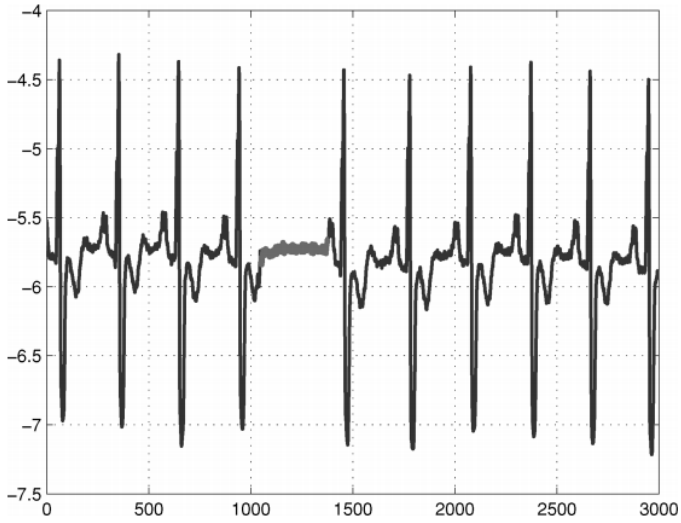
شکل ۲—۲ ناهنجاری متینی، سری زمانی دمای سه سال یک منطقه جغرافیایی
پایین آمدن دما در زمان t_1(زمستان) با t_2(تابستان) برابر است ولی چون در زمان‌های مختلفی رخ داده‌اند t_1 ناهنجاری نیست ولی t_2 یک ناهنجاری متنی است.

۲-۳-۳-ناهنجاری جمعی

اگر مجموعه‌ای از نمونه داده‌ها با توجه به کل مجموعه داده‌ها غیر عادی باشد، به آن ناهنجاری جمعی گفته می‌شود. نمونه داده‌ها به صورت فردی در یک ناهنجاری جمعی ممکن است به خودی خود ناهنجاری نباشد ، اما وقوع آنها به عنوان یک مجموعه غیر عادی است. شکل ۲-۳ خروجی الکتروکاردیوگرام[19] یک انسان را نشان می‌دهد[۷]. ناحیه برجسته شده نشان دهنده ناهنجاری است زیرا برای یک بازه زمانی به صورت غیر طبیعی رخ داده است، درحالی که این مقدارهای کم به صورت فردی(نقطه‌ای) ناهنجاری محسوب نمی‌گردند.

به عنوان مثال دیگر، یک سری از فعالیت‌های زیر را در یک رایانه در نظر بگیرید: smtp-mail، http-web، smtp-mail، buffer-overflow، ssh، ftp و ... ممکن است دنباله این رویدادها یک حمله مبتنی بر وب باشد که به وسیله کنترل از راه دور، داده‌ها را از رایانه میزبان به یک مقصد دوردست (یک رایانه دیگر) از طریق ftp انتقال دهد، که این ناهنجاری است. اما هنگامی که همین اتفاقات در مکان‌های دیگر رخ دهند، ناهنجاری نیستند.

[19] Electrocardiogram



شکل ۲—۳ ناهنجاری جمعی در خروجی الکتروکاردیوگرام انسان است

لازم به ذکر است در حالی که ناهنجاری‌های نقطه‌ای می‌تواند در هر مجموعه داده‌ای رخ دهند، ناهنجاری‌های جمعی تنها در مجموعه داده‌های رخ می‌دهند که در آن داده‌ها مرتبط باشند. در مقابل، وقوع ناهنجاری‌های متنی به در دسترس بودن ویژگی‌های بافتی در داده‌ها بستگی دارد.یک ناهنجاری نقطه‌ای یا یک ناهنجاری جمعی نیز می‌تواند یک ناهنجاری متنی باشد اگر با توجه به یک زمینه مورد تجزیه و تحلیل قرار گیرد. از این رو، مساله تشخیص ناهنجاری یک نقطه و یا مساله تشخیص ناهنجاری جمعی را می‌توان به مشکل تشخیص ناهنجاری متنی، با گنجاندن اطلاعات متنی، تبدیل کرد [۴].

۲-۴- کاربردهای ناهنجاری

تشخیص ناهنجاری در حوزه‌های مختلف تحقیقاتی و حوزه‌های متفاوت شامل بانک، تقلب، خسارت صنعتی، پردازش تصویر، بیمه، سیستم های حیاتی، مراقبت از سلامتی، نظامی و شبکه‌های اجتماعی کاربرد دارد که ما چند مورد از کاربردهای تشخیص ناهنجاری را مورد بررسی قرار می دهیم [۸]. مقالات مروری متعددی در رابطه با تشخیص ناهنجاری، تاکنون منتشر شده‌اند [۹]، [۱۰]، [۱۱] و [۱۲]. نمودار زیر به شیوه‌های دسته‌بندی مسایل مرتبط با ناهنجاری اشاره دارد:

- **تشخیص تقلب[۲۰]**

تشخیص تقلب یا تشخیص کلاهبرداری به کشف فعالیت‌های مجرمانه در سازمان‌های تجاری مانند بانک‌ها، شرکت‌های کارت اعتباری، آژانس‌های بیمه، شرکت‌های تلفن همراه، سهام بورس و غیره اشاره دارد. ممکن است کاربران مخرب، مشتری واقعی سازمان باشند یا

[20] Fraud Detection

به عنوان مشتری معرفی شوند. کلاهبرداری در شرایطی اتفاق می‌افتد که این کاربران منابع تهیه شده توسط سازمان را به روشی غیر مجاز مصرف کنند. این سازمان‌ها برای جلوگیری از خسارت‌های اقتصادی می‌خواهند که فوراً چنین کلاهبرداری‌هایی را کشف کنند. فاوکت و پرووست [۲۱] [۱۳] اصطلاح نظارت بر فعالیت را به عنوان یک رویکرد کلی برای کشف تقلب در این حوزه‌ها معرفی کردند. رویکرد معمولی روش‌های تشخیص ناهنجاری، حفظ پروفایل مشتری و نظارت بر پروفایل‌ها برای تشخیص هرگونه انحراف است. برخی از کاربردهای خاص کشف تقلب عبارت‌اند از:

۱. تشخیص تقلب در کارت‌های اعتباری [۲۲]
۲. تشخیص تقلب در گوشی‌ها موبایل [۲۳]
۳. تشخیص تقلب در مطالبات بیمه [۲۴]

- **تشخیص ناهنجاری پزشکی و بهداشت عمومی** [۲۵]

تشخیص ناهنجاری در حوزه‌های پزشکی و بهداشت عمومی به طور معمول با پرونده بیمار انجام می‌شود. این داده‌ها به دلایل مختلفی مانند وضعیت غیر طبیعی بیمار، خطاهای ابزار دقیق [۲۶] یا خطاهای ضبط [۲۷]، می‌توانند ناهنجاری داشته باشند. برای همین چندین روش در تشخیص شیوع بیماری در یک ناحیه خاص استفاده می‌شود [۱۴]. بنابراین تشخیص ناهنجاری یک مشکل بسیار مهم در این حوزه است و به درجه بالایی از دقت نیاز دارد. داده‌ها به طور معمول از پرونده‌هایی تشکیل شده است که ممکن است دارای چندین نوع ویژگی مختلف از جمله سن بیمار، گروه خونی و وزن باشد. داده‌ها همچنین ممکن است جنبه زمانی و مکانی نیز داشته باشند. بیشتر روش‌های تشخیص ناهنجاری فعلی در این حوزه با هدف شناسایی سوابق غیر عادی (ناهنجاری‌های نقطه) انجام شده است. معمولاً داده‌های دارای برچسب متعلق به بیماران سالم است، به همین دلیل بسیاری از روش‌ها یک رویکرد نیمه نظارت شده را اتخاذ کرده‌اند. چالش برانگیزترین جنبه مشکل تشخیص ناهنجاری در این حوزه این است که هزینه طبقه‌بندی ناهنجاری به صورت عادی می‌تواند بسیار زیاد باشد.

---

[21] Fawcett And Provost

[22] Credit Card Fraud Detection

[23] Mobile Phone Fraud Detection

[24] Insurance Claim Fraud Detection

[25] Medical And Public Health Anomaly Detection

[26] Instrumentation Errors

[27] Recording Errors

- **دیگر کاربردها**

مثال‌هایی شامل نفوذ در شبکه، خرابی شبکه [۱۵]، [۱۶]، [۱۷] تقلب[۲۸] در کارت‌های اعتباری [۱۷] تقلب در بیمه فرد [۱۸] تقلب در مطالبه بیمه سلامت [۱۹] بازدهی پایین حسابداری [۲۰] ایمیل‌های ناخواسته [۲۱] رای فریبکارانه و بازبینی‌های غلط [۲۲] تقلب در مزایده‌ها و حراج‌ها [۲۳] فرار از مالیات [۲۴] مشاهده فعالیت مشتریان و صفحات شخصی کاربران [۲۵] تقلب در تعداد کلیک‌ها [۲۶] و غیره همه با روش‌های تشخیص ناهنجاری سروکار دارند [۲۷].

۲-۵- روش‌های تشخیص ناهنجاری

معمولا برچسب داده‌ها نشان می‌دهد که آیا این نمونه طبیعی است یا ناهنجار. لازم به ذکر است که به دست آوردن داده‌های دارای برچسب که دقیق باشد و همچنین نماینده انواع رفتارها باشد، غالباً بسیار گران قیمت است. برچسب زدن اغلب توسط یک متخصص انسانی به صورت دستی انجام می‌شود و از این رو برای به دست آوردن مجموعه داده‌های آموزشی دارای برچسب، تلاش جدی انجام می‌شود. به طور معمول، دریافت یک مجموعه داده دارای برچسب از داده‌های غیر عادی که شامل انواع ممکن رفتارهای غیرعادی باشد دشوارتر از دریافت برچسب‌ها برای رفتارهای عادی است. علاوه بر این، رفتار ناهنجار در طبیعت اغلب پویا است، به عنوان مثال، ممکن است انواع جدیدی از ناهنجاری‌ها به وجود بیاید، که برای آن‌ها هیچگونه برچسب آموزشی وجود ندارد. بر اساس میزان دسترسی به برچسب‌ها، روش‌های تشخیص ناهنجاری به سه دسته کلی تشخیص ناهنجاری بانظارت[۲۹]، تشخیص ناهنجاری نیمه نظارتی[۳۰] و تشخیص ناهنجاری بدون نظارت[۳۱] تقسیم بندی می‌شود.

۲-۶- تشخیص ناهنجاری در گراف

یک گراف شامل مجموعه‌ی از رئوس یا گره‌ها است. گراف را با نماد $G = (V. E)$ نمایش می‌دهند که در آن V نماد مجموعه‌ی از گره‌ها و E نماد مجموعه‌ای از یال‌ها است. در یک ناهنجاری مبتنی بر گراف چندین تغییر ممکن است اتفاق بیافت [2]:

۱. یک گره غیر منتظره وجود دارد

۲. یک یال غیر منتظره وجود دارد

---

[28] Fraud

[29] Supervised Anomaly Detection

[30] Semisupervised Anomaly Detection

[31] Unsupervised Anomaly Detection

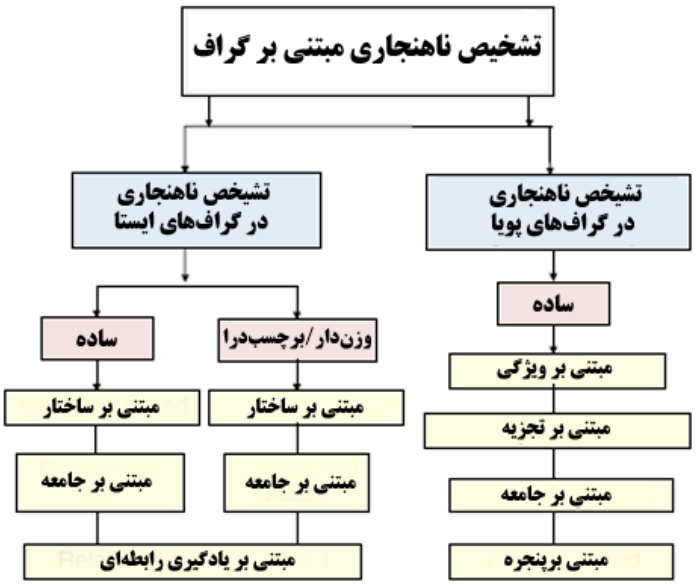
۳. یک برچسب متفات از انتظار روی گره وجود دارد

۴. یک برچسب متفاوت از انتظار روی یال وجود دارد

۵. یک گره مورد انتظار وجود ندارد(حذف شده)

۶. یک یال مورد انتظار بین دو گره وجود ندارد(حذف شده)

پیدا کردن یک شی (گره، یال و زیر ساختار) نادر در یک گراف که خیلی متفاوت‌تر از تمام شیءهای گراف است را تشخیص ناهنجاری گراف می‌گویند به عنوان مثال رکورد، نقطه، شئ گراف، به عنوان ناهنجاری در نظر گرفته می‌شود اگر بیشتر از مقدار آستانه‌ای[۳۲] باشد که کاربر تعریف کرده است. تشخیص ناهنجاری در گراف به دو دسته تشخیص ناهنجاری در گراف‌های ایستا[۳۳] و تشخیص ناهنجاری در گراف‌های پویا[۳۴] تقسیم‌بندی می‌شود [۲۷].



شکل ۲—۴ روش‌های تشخیص ناهنجاری در گراف

[32] Threshold

[33] Anomaly Detection in Static Graphs

[34] Anomaly Detection in Dynamic Graphs

۲-۶-۱- تشخیص ناهنجاری در گراف‌های ایستا

شکل ۲—۵ گراف ایستا

در این بخش به تشخیص ناهنجاری در گراف‌های ایستا می‌پردازیم. یعنی وظیفه اصلی در اینجا مشخص کردن موجودیت‌های ناهنجار و غیر عادی شبکه (به عنوان مثال گره‌ها، یال‌ها و زیرگراف‌ها) با توجه به کل ساختار گراف است. تشخیص ناهنجاری در داده‌های گراف به دو دسته زیر تقسیم می‌شوند:

۱. گراف ساده(بدون برچسب)[۳۵]

۲. گراف وابسته(برچسب‌دار)[۳۶]

یک گراف ساده شامل تنها گره‌ها و یال‌های بین گره‌ها است، یعنی از ساختار گراف تشکیل شده‌است. گراف وابسته گرافی است که گره‌ها یا لبه‌ها دارای ویژگی‌های مرتبط با هم هستند. به عنوان مثال در یک شبکه اجتماعی، کاربران ممکن است علایق مختلفی داشته باشند، مثلا در مکان‌های مختلفی کار کنند یا در مکان‌های مختلف زندگی کنند، سطوح مختلف تحصیلی و غیره دارند. در حالی که تعریف خاص ناهنجاری‌های گراف ممکن است متفاوت باشد، تعریف کلی برای مسئله تشخیص ناهنجاری برای گراف‌های ایستا به شرح زیر می‌باشد:

با توجه به یک پایگاه داده گراف (ساده یا وابسته) پیدا کردن گره‌ها، یال‌ها و زیر ساخت‌هایی که (نادر[۳۷] و متفاوت[۳۸]) به طور قابل توجهی از الگوهای مشاهده شده در نمودار منحرف شده‌اند را تشخیص ناهنجاری گراف می گویند.

---

35 Plain (Unlabeled) Graphs

36 Attributed (Node-/Edge-Labeled) Graphs

37 Few

38 Different

۲-۶-۱-۱- ناهنجاری در گراف‌های ایستا ساده

برای یک گراف ساده، تنها اطلاعات در مورد ساختار گراف لازم است. این دسته از روش های تشخیص ناهنجاری از ساختار گراف برای یافتن الگوها و ناهنجاری‌های استفاده می کنند. که خود به دو دسته الگوهای مبتنی بر ساختار [۳۹] و الگوهای مبتنی بر جامعه [۴۰] تقسیم می‌شوند.

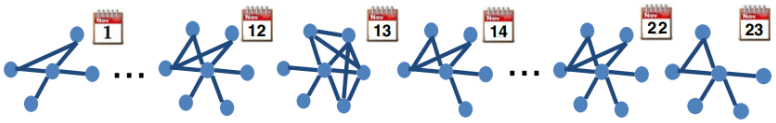
- **الگوهای مبتنی بر ساختار**

روش مبتنی بر ساختار به دو دسته مبتنی بر ویژگی[۴۱] و مبتنی بر مجاورت[۴۲] تقسیم شده است. گروه اول از ساختار گراف برای استخراج گراف مرکزی[۴۳]، مانند درجه گره و مرکزیت زیرگراف بهره می‌برد ، در حالی که گروه دوم از ساختار گراف برای تعیین کمیت نزدیکی گره‌ها در گراف برای شناسایی ارتباطات استفاده می کنند.

- **روش مبتنی بر جامعه**

روش‌های مبتنی بر جامعه برای تشخیص ناهنجاری گره در گراف استفاده می‌شود که به آن تشخیص ناهنجار در جوامع نیز گفته می‌شود. به عبارت دیگر مقدار یا ارزش صفت‌های یک گره از سایر گره‌ها منحرف شود، به عنوان مثال اگر یک فرد سیگار در تیم یا جامعه بازیکنان بیس بال وجود داشته باشد، آن ناهنجاری خواهد بود.

۲-۶-۲- تشخیص ناهنجاری در گراف‌های پویا



شکل ۲—۶ گراف پویا

39 Structure-Based Patterns

40 Community-Based Patterns

41 Feature-Based

42 Proximity Based

43 Graph-Centric

تشخیص ناهنجاری گراف‌های پویا، به عنوان تشخیص الگوی ناهنجاری زمانی[۴۴]، تشخیص رویداد[۴۵]، تشخیص نقطه تغییر[۴۶] نیز شناخته می‌شود و معمولاً به شرح زیر تعریف می‌شود، با توجه به دنباله‌ای از گراف‌های (ساده یا وابسته)، پیدا کردن یک تغییر یا یک رویداد در یک زمان[۴۷] و همچنین پیدا کردن تعداد گره‌ها و یال‌های که بیشترین تغییر را کرده‌اند [۲۷]. تشخیص ناهنجاری در گراف‌های پویا بر اساس نوع الگوریتم به چهار دسته مبتنی بر ویژگی[۴۸]، مبتنی بر تجزیه[۴۹]، مبتنی بر جامعه یا خوشه‌بندی[۵۰] و مبتنی بر پنجره[۵۱] تقسیم می‌شوند که هر روش به صورت خلاصه در زیر توضیح داده شده است. بسته به حوزه کاربرد، ویژگی‌های مورد نیاز الگوریتم‌ها متفاوت است، اما معمول ترین ویژگی‌ها مورد نظر عبارتند از [۲۷]:

۱. مقیاس پذیر[۵۲] باشد، یعنی روی انواع گراف‌های پویا با اندازه‌های مختلف جوابگو باشد.

۲. حساس به تغییر ساختاری و متنی یا زمینه‌ای[۵۳]، یعنی روش‌های تشخیص ناهنجاری باید بتوانند تفاوت‌های ساختاری بین گراف‌های ورودی تحت مقایسه (به عنوان مثال، یال‌های گمشده یا یال‌های جدید، گره‌های گمشده یا جدید، تغییر در وزن یال‌ها) و همچنین تغییر در سایر خصوصیات نمودارها، مانند برچسب‌ها، گره‌ها و یال‌ها را پیدا کنند.

۳. آگاهی از تغییرات[۵۴]، الگوریتم‌ها باید به نوع و میزان تغییر حساس باشند. تغییر در گره‌ها، لبه‌ها یا سایر خصوصیات گراف باید مهم‌تر از تغییر در ساختارهای کم اهمیت باشد.

۲-۶-۲-۱- مبتنی بر ویژگی

ایده اصلی روش های مبتنی بر ویژگی این است که گراف‌های مشابه احتمالاً دارای ویژگی‌های خاصی مانند توزیع درجه[۵۵]، قطر[۵۶]،

---

44 Temporal Anomalous Pattern Detection

45 Event Detection

46 Change-Point Detection

47 Timestamps

48 Feature Based

49 decomposition-based

50 Community Or Clustering-Based

51 window-based

52 Scalability

53 Sensitivity To Structural And Contextual Changes

54 Importance-Of-Change Awareness

55 Degree Distribution

56 Diameter

مقادیر ویژه[۵۷] هستند. رویکرد کلی در تشخیص ناهنجاری در تکامل گراف‌های پویا را می‌تواند در مراحل زیر خلاصه کرد:

۱. یک خلاصه خوب از گراف ورودی در زمان فعلی استخراج کنید

۲. گراف‌های متوالی را با استفاده از یک تابع فاصله یا تابع شباهت ، مقایسه کنید.

۳. وقتی فاصله از آستانه[۵۸] مشخص که به صورت دستی یا اتوماتیک تعریف شده، بزرگ‌تر است (یا برعکس ، اگر فاصله از آستانه کوچک‌تر است)، گراف را غیر عادی توصیف کنید.

هنگام مقایسه گراف‌های متوالی، پاسخ مشخصی در مورد ویژگی‌های گراف وجود ندارد که باید در بین زمان‌های مختلف مقایسه کرد. هر الگوریتم پیشنهادی جدید برای ساختن خلاصه گراف از تابع فاصله یا تابع شباهت استفاده می‌کند همچنین از نحوه تعیین و انتخاب آستانه برای پرچسب گذاری نمونه داده به عنوان ناهنجاری استفاده می‌کند.

۲-۶-۲-۲- مبتنی بر تجزیه

ایده اصلی رویکردهای مبتنی بر تجزیه، تشخیص ناهنجاری‌های زمانی توسط تجزیه ماتریس یا تانسور در گراف‌های است که در حال تحول زمانی هستند. این روش به دو دسته ماتریس و تنسور از نظر نمایش تقسیم می‌شوند. دسته اول، روش ماتریسگرا از ویژگی‌های گرافی تولید شده توسط SVD، تجزیه عدد خاص یا NMF استفاده می‌کند. بنابراین می‌توان آن را به عنوان روش‌های تشخیص ناهنجاری تجزیه شده دسته بندی کرد. دسته دوم روش‌های تشخیص رویداد مبتنی بر تجزیه است که از تانسورها به جای ماتریس برای نمایش گراف استفاده می‌کند.

۲-۶-۲-۳- مبتنی بر جامعه

ایده اصلی روش‌های مبتنی بر جامعه یا رویکردهای مبتنی بر خوشه بندی، به این صورت است که به جای نظارت بر تغییرات در کل شبکه، خوشه یا جامعه گراف را مانیتور می‌کند و هنگام تغییر ساختاری یا تغییر متنی گزارش وقایع را تولید می‌کند.

۲-۶-۲-۴- مبتنی بر پنجره

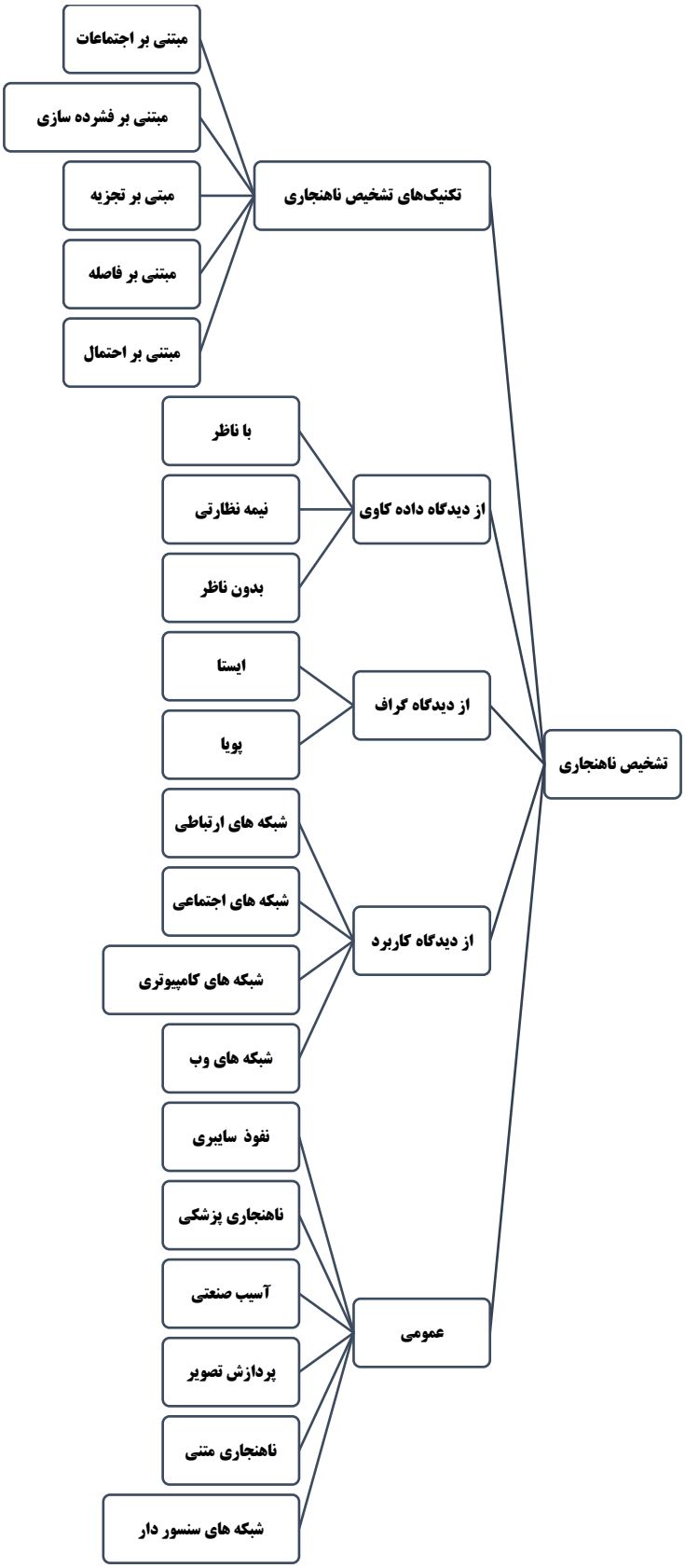
ایده اصلی آخرین دسته الگوریتم‌های تشخیص ناهنجاری گراف شامل روش‌هایی است که در یک زمان فقط به یک پنجره برای تشخیص ناهنجاری نقطه‌ای و رفتارهای ناهنجار محدود می‌شوند. در اصل، از یک تعداد نمونه قبلی برای مدل کردن رفتار "عادی" استفاده می‌شود. گراف ورودی با ویژگی‌های عادی و ناهنجار مقایسه می‌شوند. در شکل ۲-۷ دسته‌بندی و روش‌های تشخیص ناهنجاری ناهنجاری در

---

[57] Eigenvalues

[58] threshold

حوزه‌های مختلف ارائه شده است.



شکل ۲—۷ دسته‌بندی‌های مختلف از تشخیص ناهنجاری

۲-۷- مروری بر روش‌های تشخیص ناهنجاری یال در گراف

در این بخش مروری بر روش‌های مهم و اخیر تشخیص ناهنجاری یال در گراف انجام خواهد شد. یال‌های ناهنجار در روش‌های مبتنی بر ساختار به این صورت تعریف می‌شود: یال‌های که بین نودهای ناهنجار قرار داردند، ناهنجار هستند به عبارت دیگر ارتباطات بین گره‌های با درجه پایین ناهنجار محصوب می‌شوند. در [۲۸] روشی بدون نظارت برای تشخیص یال‌های جالب یا مسیرهای جالب مبتنی بر معیار کوتاه‌ترین مسیر ارائه داده است که می‌توان آن را بر روی مجموعه داده‌های چند رابطه‌ای [۵۹]اجرا کرد. این روش به اطلاعات، الگوهای موجود یا یادگیری متکی نیست و می‌تواند یال‌های جالب و جدید را قبل از تجزیه و تحلیل تشخص بدهد. در [۲۹] یال‌های ناهنجار به شبکه اضافه می‌کنند و معیار کتز را برای آن محاسبه می‌کنند، اگر معیار کتز برای یال پایین باشد یعنی یال ناهنجار است. از این روش برای پیدا کردن رابطه‌های ناهنجار بین شبکه‌ی علمی پژوهشگران استفاده شده است. در [۳۰] ر وشی برای پیدا کردن یال‌های ناهنجار و یال‌های گم شده با استفاده از یک چهار چوب ریاضی ارائه شده است. در[۳۱] سه روش معیار بینابینی یال [۶۰]، معیار مرکزیت لود یال [۶۱] و معیار گام‌برداری تصادفی بینابینی یال [۶۲] برای تشخیص یال‌های ناهنجار ارائه شده است. در [۳۲] دو روش MIDAS و MIDAS-R مبتنی بر میکرو کلاستر [۶۳] را برای تشخیص یال‌های ناهنجار به صورت استریم [۶۴] ارائه شده است.

در روش‌های که در ادامه معرفی می‌شود از لاین گراف برای تغییر حالت به تشخیص ناهنجاری یال استفاده شده است(در توضیحات تکمیلی آخر همین فصل کامل توضیح داده شده است). در مقاله [۳۳] الگوریتم تجریه پیازی [۶۵] که به آن k-core Decomposition هم گفته می‌شود، شبکه را با هدف تبدیل به هسته‌های تو در تو هرس و هسته و حاشیه هسته را مشخص می‌کند. الگوریتم تجزیه پیازی به این صورت عمل می‌کند: اول گره‌های با در جه ۱ را حذف می‌کند، بعد از حذف گره‌های با درجه یک، دوباره درجه همه گره‌ها را محاسبه می‌کند و اگر گره‌های با درجه یک دوباره پیدا شود آن‌ها را حذف می‌کند تا زمانی که همه گره‌های درجه یک حذف شوند این کار بصورت تکراری ادامه پیدا می‌کند، که این اولین پوسته پیاز است. بعد به k یک واحد اضافه می‌کند و همان کار در این مرحله تکرار می‌شود یعنی گره‌های با درجه ۲ حذف می‌شوند و این کار تا زمانی که همه گره‌های درجه ۲ حذف نشده‌اند ادامه پیدا می‌کند. این روال تا جایی ادامه پیدا می‌کند که هر گره‌ای

---

[59] Multi-Relational Datasets

[60] Edge Betweenness Centralities

[61] Edge Load Centralities

[62] Edge Centralities

[63] Micro Cluster Based

[64] Streams

[65] Onion Decomposition

با هر درجه‌ای حذف شود. در مقاله ۲۰۱۸ که [۳۴] [۶۶]، در مورد معیارهای مرکزیت گره‌ها است که همه معیارهای مرکزیت را معرفی کرده است. ما چهار معیار معروف مرکزیت را به طور خلاصه معرفی و در این پایان نامه استفاده خواهیم کرد. معیار اول مرکزیت درجه[۶۷] درجه یک گره تعداد گره هایی است که با آن گره در همسایگی مستقیم قرار دارد. هر چقدر درجه یک گره بیشتر باشد، اهمیت آن گره بیشتر می شود. مثلا در شبکه وب، اگر درجه ورودی یک دامنه زیاد باشد، یعنی سایت مرجع است و اگر درجه خروجی آن زیاد باشد، یعنی اینکه سایتی است که از اطلاعات (اخبار) سایت‌های دیگر استفاده می کند. معیار مرکزیت دوم، معیار مرکزیت بینابینی[۶۸] است. بینابینی عبارت است از نسبت تعداد دفعاتی که یک گره یا یک یال بر روی کوتاهترین مسیر میان نودهای مختلف یک گراف قرار می گیرد. بینابینی یک نود خاص در شبکه عبارت است از تعداد کوتاهترین مسیرهای میان نودهای شبکه که از یک نود خاص رد می‌شود. در حقیقت این معیار محاسبه می کند چه تعداد از نودهای شبکه برای ارتباط سریعتر با هم (با واسطه کمتر) به این نود نیاز دارند. هر چه بینابینی نود زیادتر باشد یعنی اینکه نود در مکان استراتژیک‌تری قرار گرفته است. همچنین نشان دهنده درصدی از اطلاعات است که از یک گره می گذرد و مشخص کننده توانایی یک گره برای تسهیل گسترش ارتباط بین سایر عناصر گره های گراف است و در واقع نمایشی برای میزان قابلیت هر گره برای کمک به دسترسی سایرین به اطلاعات و یا گسترش یک تاثیر در شبکه می‌باشد. این معیار برای یافتن محل افرادی که توانایی مرتبط ساختن با جفت‌ها و گروه‌های دیگر را دارند، می‌باشد. نزدیکی[۶۹] عبارت از عکس متوسط فاصله یک گره تا گره‌های دیگر گراف می‌باشد. گره‌ای که دارای بیشترین مقدار نزدیکی است سرعت دسترسی بیشتری به گره‌های دیگر دارد و می تواند در مدت زمان کمی به همه نودها اطلاعات ارسال نماید یا از آنها اطلاعات بگیرد. روش مرکزیت مقادیر ویژه [۷۰]، [۴۲] اهمیت گره‌ها را بر اساس گره‌های مجاور محاسبه می کند. این محاسبه در گراف‌های با اتصال قوی اتفاق می‌افتد. اگر گره‌ای به گره‌هایی که دارای اهمیت بالایی هستند متصل باشد تحت تأثیر آن‌ها اهمیت او نیز بالا می‌رود. این روش به‌صورت تکراری برای محاسبه گره اهمیت همسایگان را نیز در نظر می گیرد. ابتدا به همه گره‌ها یک امتیاز اولیه داده می‌شود. در ادامه به‌صورت زنجیره‌ای تا زمانی که به پایداری برسد این امتیازدهی ادامه می‌یابد. امتیازدهی در این روش بر اساس این مفهوم است که گره‌های با اتصالات بالا به گره‌های دنبال کننده آن‌ها از نظر امتیاز کمک می کنند. روش رتبه‌بندی صفحه ازاین‌روش الگوبرداری کرده است.

---

[66] Survey

[67] Degree Centrality

[68] Betweenness Centrality

[69] Closeness

[70] Eigenvector Centrality

۲-۸- پیشگویی پیوند[71]

اخیرا پیشگویی پیوند مورد توجه محققان زیادی در حوزه‌های مختلف قرار گرفته است. پیشگویی پیوند به پیشگویی یال بین دو گره بر اساس ویژگی‌های گره‌ها و ویژگی‌های کلی شبکه می‌پردازد. این بحث کاربردهای مهمی در زمینه‌های مختلف مانند شبکه اجتماعی، سیستم‌های بیولوژی و سایر شبکه ها دارد. پیشگویی پیوند به طور گسترده‌ای در شبکه‌های بیولوژیکی مانند شبکه تعامل پروتئین، شبکه متابولیک و شبکه غذایی مورد استفاده قرار گرفته است[۳۵, ۳۶]. از پیشگویی پیوند برای یافتن پیوندهای مفقود شده استفاده می‌شود و در صورت دقیق بودن پیشگویی‌ها به کاهش هزینه‌های آزمایشگاهی کمک می‌کند. همچنین در شبکه‌های تعاملی، شبکه‌های علمی و تجاری می‌تواند نقش مهمی در پیشگویی انجمن‌های جدید داشته باشد[۳۷]. به علاوه این یکی دیگر از کاربردهای پیشگویی پیوند در زمینه سیستم‌های توصیه‌گر است: خدماتی که تقریبا توسط همه شبکه‌های اجتماعی ارائه می‌شود و بیشتر در تجارت الکترونیکی مورد استفاده قرار گرفته است[۳۸]. پیش بینی لینک همچنین می تواند در یافتن پیوندهای پنهان در شبکه های جنایی مفید باشد که یکی دیگر از زمینه‌های مهم پژوهشی است[۳۹].

پیشگویی پیوند اساساً می‌تواند دو نوع باشد: ساختاری[72] و زمانی[73] که در شکل ۲-۸ نشان داده شده است.

پیشگویی پیوند ساختاری به مشکل یافتن پیوندهای مفقود یا پنهان که احتمالاً در یک شبکه وجود دارند، اشاره دارد. با استفاده از داده های قابل مشاهده شبکه این امر بر وجود پیوندهایی که به طور مستقیم قابل مشاهده نیستند، تمرکز دارد. این روش برای یافتن الگوهای پنهان ژن‌ها ، تداخلات پروتئین برای مطالعات پزشکی در مورد بیماری‌های مختلف مانند سرطان، ایدز، آلزایمر و... کاربرد دارد[۳۵, ۳۶, ۳۹].

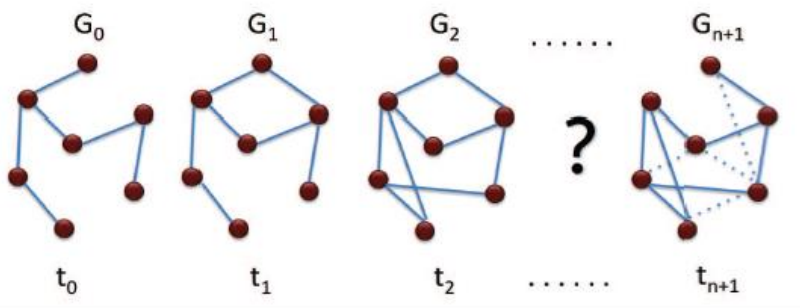
پیشگویی پیوند زمانی به مشکل یافتن لینک های جدید با مطالعه تاریخچه زمانی یک شبکه، اشاره دارد. در این روش در مورد شبکه تا زمان t اطلاعات داریم و هدف پیشگویی، پیوند جدیدی است که ممکن است در مقطعی از زمان (t+k) در آینده به وجود بیاید. این نوع پیشگویی پیوند به طور گسترده در سیستم‌های توصیه‌گر شبکه‌های اجتماعی برای پیشنهاد دوست، در وب سایت‌های تجاری برای محصولات و پیشنهاد عبارت‌های کلیدی در موتورهای جستوگر کاربرد دارد[۳۷, ۳۸, ۴۰].

---

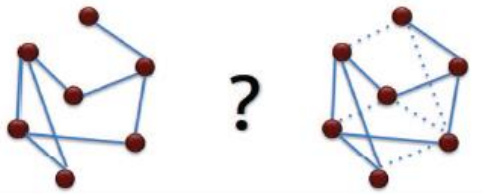
71 Link Prediction

72 Structural link prediction

73 Temporal link prediction



الف) پیشگویی پیوند زمانی: پیدا کردن پیوند جدید



ب) پیشگویی پیوند ساختاری: پیدا کردن پیوند مخفی شده یا گم شده

شکل ۲—۸انواع پیشگویی پیوند

پیشگویی وجود ارتباط میان دو موجودیت بر اساس ویژگی‌های موجودیت‌ها و دیگر پیوندهای مشاهده شده در گراف را پیشگویی پیوند می‌گویند[۴۱]. یا به عبارت دیگر اگر در زمان t یک تصویر لحظه‌ای از مجموعه پیوندها داشته باشیم، هدف پیشگویی پیوندها در زمان $t+1$ است. در مقاله[۴۰] پیشگویی پیوند را به این صورت تعریف کرده است: فرض کنید یک شبکه اجتماعی به صورت $G = (V.E)$ داشته باشیم، به طوری که هر یال $e = (u.v) \in E$ نمایانگر یک تعامل میان u و v در زمان $t(e)$ باشد. تعامل میان u و v در زمان‌های مختلف به صورت یال‌های موازی در درون گراف نشان داده می‌شوند. در دو زمان $t < t'$ که به صورت $G[t.t']$ نیز نمایش داده می‌شود، زیر گرافی از G است که شامل تمامی یال‌ها در بازه زمانی t و t' است. اکنون تعریف پیش گویی پیوند به این صورت بیان می‌شود که: با انتخاب چهار بازه زمانی $t_0 < t'_0 < t_1 < t'_1$ الگوریتم پیشگویی پیوند تنها با دسترسی به گراف $G[t_0.t'_0]$ باید در خروجی یال‌هایی را پیشگویی کند که در گراف $G[t_0.t'_0]$ وجود ندارند ولی در گراف $G[t_1.t'_1]$ وجود داشته باشند. در پیشگویی پیوند ما یک عکس فوری از یک شبکه داریم و می‌خواهیم براساس تعاملات میان اعضای موجود در شبکه پی ببریم که به احتمال زیاد چه ارتباط جدیدی رخ می‌دهد یا در آیندی نزدیک چه تعاملاتی از بین می‌رود [۴۰]. اگر چه این موضوع به طور گسترده مورد مطالعه قرار گرفته، ولی چالش‌های زیادی در انجام و چگونگی این موضوع وجود دارد.

۲-۹- روش‌های پیشگویی پیوند

می‌توان روش‌های مختلف پیشگویی پیوند را به سه دسته کلی بدون نظارت، بانظارت و نیمه نظارتی تقسیم بندی کرد که هر دسته چندین زیر مجموعه دارد. روش‌های زیادی برای پیشگویی پیوند بدون نظارت پیشنهاد شده است این دسته از روش‌های پیشگویی پیوند نیازی به داده‌های آموزشی یا داده‌های برچسب خورده ندارند. اما روش‌های پیشگویی پیوند با نظارت به داده‌ها آموزشی نیاز دارند و برای این دسته هم روش‌های زیادی پیشنهاد شده است. در روش‌های پیشگویی پیوند نیمه نظارتی تا حدودی به داده‌های آموزشی نیاز داریم برای این دسته روش‌های خیلی کمی ارائه شده است [۴۲].

۲-۹-۱- پیشگویی پیوند بدون نظارت

برای پیشگویی پیوند، روش‌های بدون نظارت زیادی وجود دارد که در بر اساس ساختار شبکه، به پیوند بین جفت گره‌ها امتیاز داده می‌دهند. این امتیازات شباهت بین دو گره‌ها را نشان می‌دهند که احتمال ایجاد ارتباط یا پیوند بین آنها است.

۲-۹-۱-۱- روش‌های مبتنی بر همسایگی

۱) روش همسایه‌های مشترک[۷۴]

در مقاله بیان شده است که احتمال ایجاد لینک‌هایی در آینده در یک شبکه، با تعداد مجاوران مشترک ارتباط مثبت دارد. معمولا دو راس که تعداد همسایه‌های مشترک زیادی دارند به این معنی است که احتمال اینکه در آینده با هم همکاری کنند، بیشتر است. اگر $\Gamma(X)$ نشان دهنده همسایه‌های گره x باشد همسایه‌های مشترک دو گره از فرمول زیر بدست می آید.

$$Cn(x.y) = |\Gamma(x) \cap \Gamma(y)| \qquad (۱-۲)$$

اگر گراف وزندار باشد از رابطه زیر برای پیش‌گویی پیوند استفاده می‌کنید.

$$CN(x.y) = \sum_{z\epsilon\Gamma(x)\cap\Gamma(y)} w(x.z) + w(y.z) \qquad (۲-۲)$$

۲) روش جاکارد[۷۵]

این الگوریتم با نرمال کردن فرمول الگوریتم همسایه‌های مشترک به دست می‌آید. یعنی نسبت تعداد دوستان مشترک دو

---

[74] Common neighbors

[75] Jaccard

گره(همسایه‌های مشترک) به مجموع همسایه‌های آن دو تعریف می‌شود. این معیار هرچه به یک نزدیک‌تر باشد، نشان‌دهنده شباهت بیشتر دو گره مورد بررسی است. این سنجش روشی است برای شناسایی محتوای مشترک که در بازیابی اطلاعات هم کاربرد دارد.

$$JC(x.y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \qquad (۳-۲)$$

اگر گراف وزن‌دار باشد به صورت زیر است:

$$JC(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) + w(y,z)}{\sum_{a \in \Gamma(x)} w(x,a) + \sum_{b \in \Gamma(y)} w(y,b)} \qquad (۴-۲)$$

۳) روش آدامیک-آدار[۷۶]

تعداد همسایه‌های مشترک دو راس را محاسبه می‌کند با این تفاوت که به هر همسایه مشترک یک وزن براساس معکوس تعداد یال‌های متصل به آن می‌دهد. در این روش وقتی تعداد همسایه‌های یک راس کمتر باشد ارزش هریک از همسایه‌های آن بیشتر خواهد بود.

$$\sum_{s \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log_2(|\Gamma(s)|)} \qquad (۵-۲)$$

برای گراف وزن‌دار داریم:

$$AA(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) + w(y,z)}{log(1 + \sum_{c \in \Gamma(z)} w(z,c))} \qquad (۶-۲)$$

۴) روش اندیس تخصیص منبع[۷۷]

اندیس تخصیص منبع یک ویژگی بسیار خوب برای پیشگویی پیوند در شبکه است. این ویژگی شبکه را مجموعه ای از رئوس به هم پیوسته فرض می‌کند که جریان اطلاعات در آن جاری می‌شود و افراد از طریق این جریان‌ها می‌توانند به یکدیگر متصل شوند. در واقع هر فرد دارای منابع است و این منابع را در طی زمان به همسایگان خود منتقل می‌کند. در این روش شباهت دو راس براساس میزان منابع مشترکی که آن دو در شبکه دریافت می‌کنند قابل محاسبه است :

---

[76] Adamic-Adar

[77] Resource Allocation

$$RA(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{|\Gamma(z)|} \qquad (۷-۲)$$

اگر گراف وزن‌دار باشد به صورت زیر است:

$$RA(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w\,(x,z) + w\,(y,z)}{\sum_{c \in \Gamma(z)} w(z,c)} \qquad (۸-۲)$$

۵) روش پیوست امتیازی[۷۸]

این معیار بر این ایده است که کاربران با دوستان زیاد به ایجاد ارتباطات بیشتر در آینده تمایل دارند.

$$\boldsymbol{PA(x.y) = |\Gamma(x)| \cdot t|\Gamma(y)|} \qquad (۹-۲)$$

اگر گراف وزن‌دار باشد به صورت زیر است:

$$PA(x,y) = \sum_{a \in \Gamma(x)} w(x,a) * \sum_{b \in \Gamma(y)} w(y,b) \qquad (۱۰-۲)$$

۲-۹-۱-۲- روش‌های مبتنی بر مسیر

۱) روش کوتاه‌ترین مسیر[۷۹]

کم‌ترین فاصله(تعداد یال) بین گره x و y در گراف G را کوتاه‌ترین مسیر بین آن دو گره می‌گویند. هر چی مسیر طولانی‌تر باشد یعنی آن دو گره به هم شباهتی ندارند و برعکس. این حقیقت که دوستانِ دوستان می‌توانند با هم دوست شوند نشان می‌دهد که فاصله بین دو مسیر در شبکه می‌تواند شکل‌گیری ارتباط بین آن‌ها را تحت تاثیر قرار دهد.

۲) روش کاتز[۸۰]

ویژگی کاتز براساس تعداد مسیرهایی است که بین دو راس وجود دارد. اهمیت کاتز آن است که اگر بین دو راس تعداد بیشتری مسیر وجود داشته باشد احتمال این که دو راس در آینده با هم همکاری کنند، بیشتر است. چون از واسطه‌های بیشتری می‌توانند به هم برسند.

---

[78] Preferential Attachment

[79] Shortest Path Distance

[80] Katz

محاسبه این ویژگی بر اساس رابطه زیر است:

$$K(x.y) = \sum_n \beta^n path_{x.y}^n \qquad (۱۱-۲)$$

در محاسبه همان‌طور که به طور منطقی به ذهن می رسد، مسیرهای کوتاه‌تر باید ارزش بیشتری داشته باشند. به همین دلیل یک ضریب β بین صفر و یک وجود دارد که ارزش مسیرها را براساس طول مسیر تعیین می‌کند و با طولانی‌ترشدن مسیر ارزش مسیر را کاهش می‌دهد.

۳) روش قدمزنی تصادفی[۸۱]

یکی از روش‌های پیش‌گویی پیوند، استفاده از قدمزنی تصادفی برای پیدا کردن نزدیکی بین دو گره است. هر چه قدر دو گره به هم نزدیک‌تر باشد، احتمال برقراری یال در آینده بیشتر است. قدمزنی تصادفی، از یک گره مبدأ شروع به حرکت می‌کند. در هر قدم، از بین همسایه‌های گره فعلی، گرهی را با احتمال $1/K_v$ انتخاب می‌کند (K_v درجه‌ی گره فعلی است) و با آن حرکت می‌کند. سپس به همین شکل به مسیر خود ادامه می‌دهد. قدمزن تصادفی به هر گرهی که برسد یک امتیاز به آن می‌دهد. گره‌هایی که امتیاز بیشتری کسب کرده‌اند یعنی از طریق گره اول قابل دسترس‌ترند. بنابراین احتمال اینکه در آینده گره مبدأ با آن‌ها پیوند برقرار کند بیشتر است.

۲-۹-۲- پیشگویی پیوند بانظارت

این دسته از روش‌ها با یک یا چند مرحله یادگیری از فرآیند به وجود آمدن پیوندها درگذشته، به پیشگویی پیوند می‌پردازند. این دسته از روش‌ها را می‌توان روش‌های پیشگویی پیوند بانظارت نامگذاری نمود، الگوریتم‌های پیشگویی پیوند بدون نظارت غالباً از ویژگی‌های ساختاری منجمله تعداد همسایه‌های مشترک، طول کوتاهترین مسیر میان دو گره، درجه دو گره و از این قبیل ویژگی‌های ساختاری موجود در گراف شبکه‌های اجتماعی استفاده می‌نمایند. الگوریتم‌های پیشگویی پیوند بانظارت گاهی اوقات با یادگیری پارامترهای یک مدل احتمالاتی و یا بررسی روند تکامل یک زیرساختار خاص در گراف شبکه به پیشگویی پیوند می‌پردازند که به دو روش خوشه‌بندی و استخراج ویژگی دسته‌بندی می‌شود. البته پیشگویی پیوند به‌صورت بانظارت جزو تحلیل دینامیک محسوب می‌شود. اگرچه ویژگی‌های ساختاری شبکه معیار خوبی برای احتمال به وجود آمدن لینک در آینده باشد ولی بسیاری دیگر از روش‌ها هستند که می‌توانند دقت الگوریتم را بهبود ببخشند. ازآنجا که مسئله پیشگویی پیوند به حوزه‌های مختلفی مربوط است، برای آن روش‌های مختلفی در نظر گرفته‌اند. بیشتر این روشها معمولاً مبتنی بر ویژگی‌های ساختاری و روش‌های با ناظر هستند. همانطور که توضیح داده شد معمولاً این روش‌ها مبتنی بر شباهت بین دو گره استوار هستند. این روش‌ها به اطلاعات محلی شبکه وابسته هستند. بسیاری از پژوهش‌های محققین در تحلیل شبکه‌های اجتماعی روی استفاده از اطلاعات

---

[81] Random Walk

گره‌ها و رابطه بین گره‌ها مثل رابطه دوستی، اطلاعات کلاسترها و انجمن‌ها متمرکز شده است. این اطلاعات برای بهبود عملکرد پیشگویی پیوند به کار می‌روند. تجربه نشان می‌دهد که برای یک شبکه با ساختار کلاستری کم، معیارهای پیشگویی پیوند مبتنی بر ساختاری شباهت میتواند ضعیف عمل کند. در شکل ۲-۱۰ انواع روش‌های پیشگویی پیوند بانظارت ذکر شده است.

۲-۹-۳- پیشگویی پیوند نیمه نظارتی

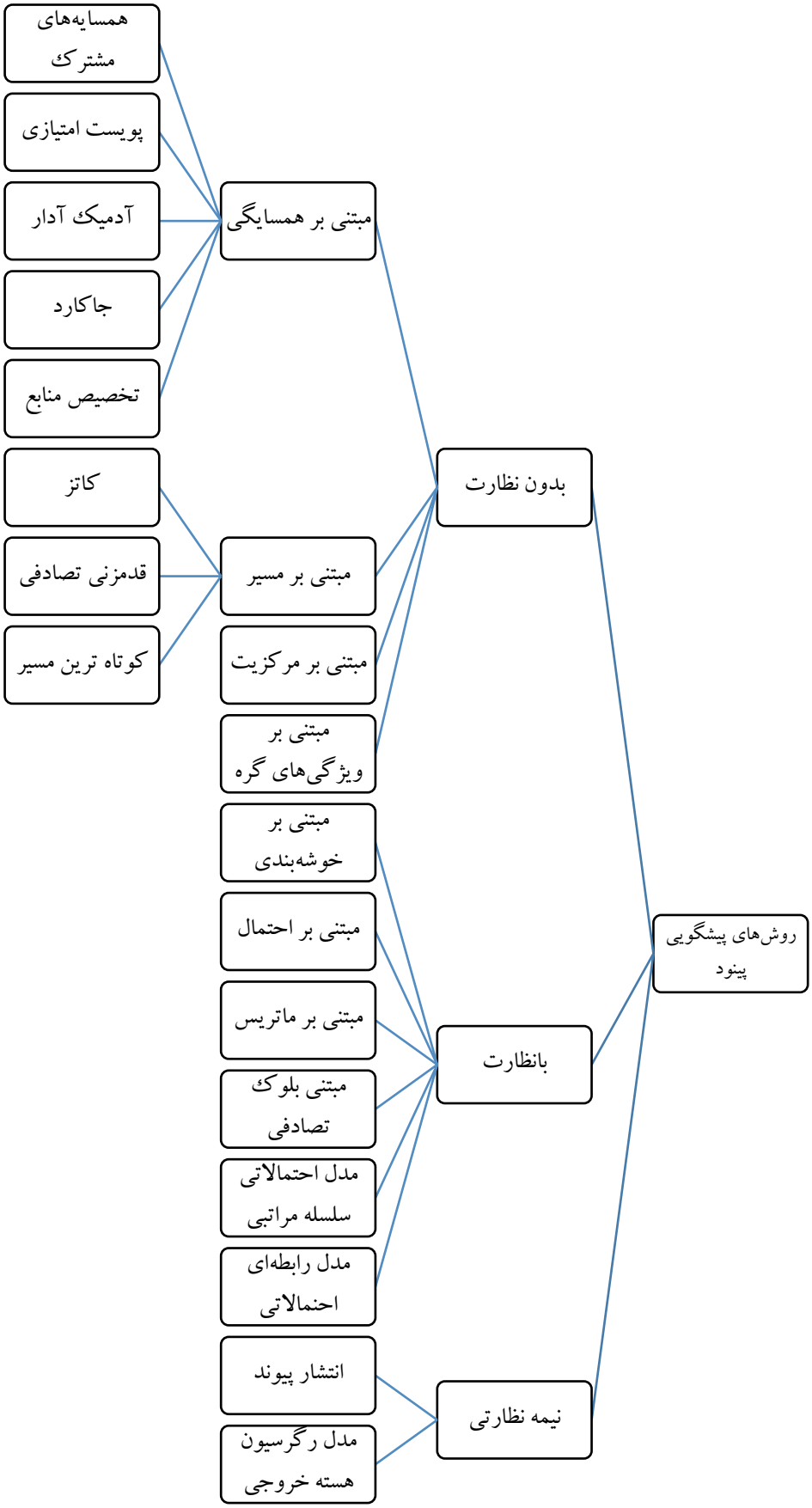
یادگیری نیمه نظارتی نوعی از یادگیری است که در آن از داده‌های بدون برچسب به همراه مقدار کمی از داده‌های برچسب‌دار استفاده می‌شود. روش‌های یادگیری بانظارت برای آموزش یک مدل (به ویژه در کارهای طبقه‌بندی) به داده‌های برچسب‌دار نیاز دارند. با این حال، به دست آوردن داده‌های برچسب خورده اغلب سخت، گران و یا وقت‌گیر هستند زیرا تهیه این نوع داده‌ها نیاز متخصص با تجربه انسانی دارد. در عین حال، جمع آوری داده‌های بدون برچسب ممکن است نسبتاً آسان باشد، اما روش‌های کمی برای استفاده از آنها وجود دارد. یادگیری نیمه نظارتی این مشکل را با استفاده از مقدار زیادی از داده‌های بدون برچسب، همراه با داده‌های برچسب‌دار، برای ساخت طبقه‌بندی بهتر برطرف می‌کند. از آنجا که یادگیری نیمه نظارتی به تلاش انسانی کمتری نیاز دارد و دقت بالایی را هم نشان می‌هد از نظر تئوری و عملی بسیار مورد توجه است[۴۳]. این نوع یادگیری در زمینه پیشگویی پیوند به خوبی مورد کاوش قرار نگرفته است و تعداد خیلی کمی از مقالات در این زمینه وجود دارد. یکی از مقاله‌های خوب در این زمینه کاشیما و همکارانشان می‌باشد[۴۴] که از مفهوم انتشار برچسب[۸۲] استفاده می‌کند. یکی دیگر از پژوهش‌های که در این زمنیه انجام شده بروارد و همکارانش است[۴۵] که مبتنی بر رگرسیون هسته خروجی[۸۳] است.

در شکل ۲-۱۰ دسته‌بندی کاملی از روش‌های پیشگویی پیوند ارائه شده است.

---

[82] Label Propagation

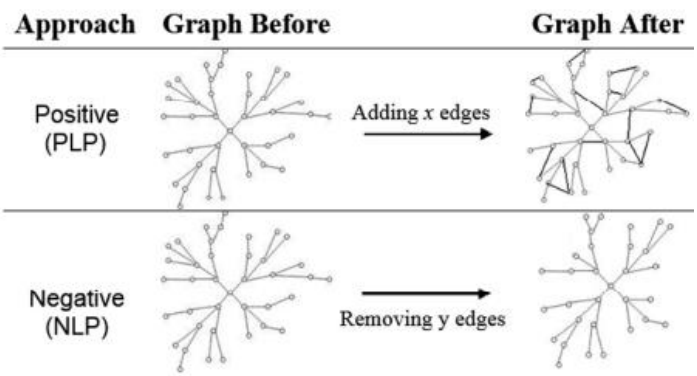
[83] Output Kernel Regression



شکل ۲—۹ دسته‌بندی روش‌های پیشگویی پیوند[۴۲]

۲-۹-۴- پیشگویی پیوند منفی[۸۴]

بیشتر روش‌های پیشگویی پیوند روش‌های پیشگویی پیوند مثبت[۸۵] هستند. پیشگویی پیوند مثبت یعنی پیشگویی پیوندهای که در آینده به وجود می‌آیند. در مقابل مقاله‌های کمی در مورد پیشگویی پیوند منفی بحث شده است. پیشگویی پیوند منفی یعنی پیشگویی پیوندهای که واقعی نیستند یا در آینده حذف می‌شوند. در مقاله [۴۶] مسئله پیشگویی پیوند را اینگونه بیان کرده است: با توجه به ساختار شبکه در زمان t_0 ، لینک‌های را پیش‌گویی می‌کنیم که در زمان $t(t < t_0)$ حذف یا ناپدید خواهند شد. پیشگویی پیوند منفی شبیه پیشگویی پیوند مثبت کار می‌کند و آن را به صورت مکمل بر روی ماتریس تقریبی بولی اعمال می‌کند. تکمیل یک ماتریس بولین با تبدیل صفر به یک و برعکس ایجاد می‌شود. لبه‌های که به وسیله ماتریس مکمل پیدا می‌شود لبه‌های هستند که به احتمال زیاد توسط پیشگویی پیوند مثبت از ماتریس تقارب واقعی حذف شده است.



شکل ۲—۱۰ پیشگویی پیوند منفی و پیشگویی پیوند مثبت

۲-۹-۵- کاربردهای پیشگویی پیوند

پیش بینی لینک برای طیف گسترده‌ای از حوزه‌ها قابل استفاده است. ما به تشریح چندتا ازکاربردهای پیشگویی پیوند در حوزه‌های مختلف می‌پردازیم.

شبکه‌های اجتماعی: علاوه بر کمک به تجزیه و تحلیل شبکه‌ها با داده های از دست رفته، الگوریتم‌های پیش‌بینی لینک می‌توانند

[84] Negative Link Prediction (NLP)

[85] Positive Link Prediction (PLP)

برای پیش بینی لینک‌هایی که درآینده‌ی شبکه‌ی در حال تکامل ایجاد می‌شوند، به کار روند، برای مثال، در شبکه‌های اجتماعی برخط، لینک‌های محتمل که هنوز وجود ندارند می‌توانند به عنوان روابط نویدبخش در یافتن دوستان جدید، به کاربران کمک کنند.

تشخیص و نفوذ شبکه‌های تروریستی: شبکه‌های اجتماعی به محیطی مناسب برای مطرح کردن اندیشه‌های گروه‌های تروریستی تبدیل شده است و توانسته‌اند هوادارانی را پیدا کنند. در این شبکه‌ها می‌توان به کمک روش‌های پیش‌بینی پیوند تلاش آن‌ها برای جذب نیرو و افرادی که به احتمال بالایی در آینده با آن‌ها همکاری خواهند داشت را پیدا نمود و آن‌ها و همچنین مجرمین را شناسایی کرد.

مسیریابی: مشخص کردن مسیرهای بهینه (ترافیک داده‌ها) با کمک پیشگویی پیوند باعث بهبود عملکرد مسیریابی در شبکه‌های مختلف مانند شبکه‌های حسگر بیسیم می‌شود.

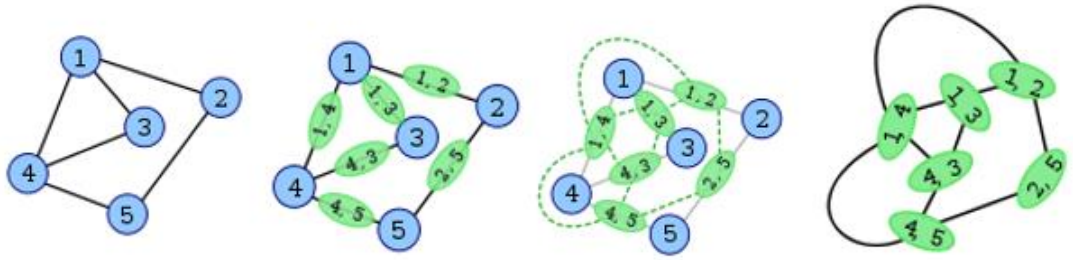
کاربردهای دیگر

۱. تراکنش‌های مالی
۲. شبکه وب(اسپم‌ها و بدافزارها)
۳. پیش‌بینی بیماری‌ها
۴. بیوانفورماتیک
۵. سیستم‌های توصیه‌گر

۲-۱۰- توضیحات تکمیلی

در بخش بررسی نتایج، روش پیشنهادی اول را با هشت الگوریتم دیگر مقایسه کرده‌ایم که از این هشت الگوریتم، چهار الگوریتم یال‌های ناهنجار را حذف و چهار الگوریتم دیگر گره‌های ناهنجار را حذف می‌کنند. ما برای این که یال‌های ناهنجار را با الگوریتم‌های تشخیص گره ناهنجار، مقایسه کنیم، از لاین گراف[۸۶][۴۷] استفاده کردیم. در لاین گراف یال‌ها را به گره و گره‌ها را به یال تبدیل می‌کنیم. لاین گراف به این صورت عمل می‌کندکه یال بین دو گره را با نام آن دو گره در نظر می‌گیرد و این کار را برای همه یال‌ها انجام می‌دهد بعد گره بین دو یال را به یال و یال‌ها را به گره تبدیل می‌کند. در شکل زیر تبدیل گراف به لاین گراف نمایش داده شده است.

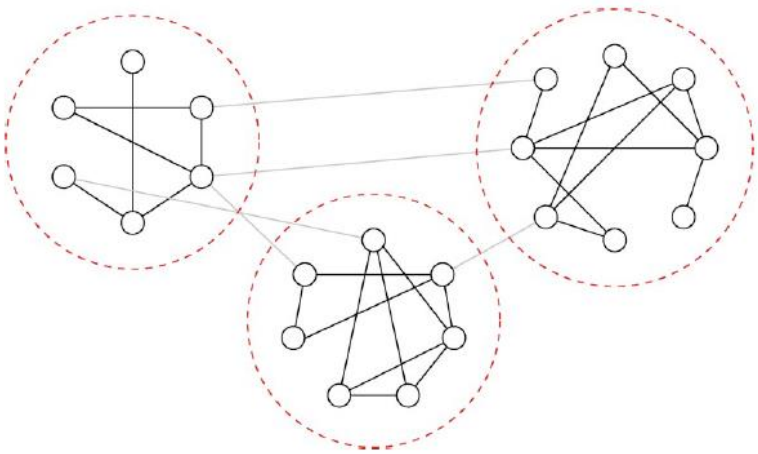
[86] Line Graph



شکل ۲—۱۱ نحوه تبدیل گراف به لاین گراف

تشخیص جوامع[87]

یک ویژگی که به نظر می‌رسد برای بسیاری از شبکه‌ها مشترک است، ساختار جامعه است. تقسیم گره‌های شبکه به گروه‌هایی که اتصالات داخل گروه‌ها متراکم اما اتصالات بین آنها پراکنده است را جامعه[88] می‌گویند مانند شکل ۴-۱ [۴۸].



شکل ۲—۱۲شبکه‌ای کوچک با ساختار جامعه

شکل ۴-۱ در این حالت، سه اجتماع وجود دارد که توسط دایره‌ها مشخص شده‌اند و دارای پیوندهای داخلی متراکم هستند، اما بین آن‌ها چگالی کمتری از یال‌های خارجی وجود دارد [۴۸].

الگوریتم‌های تشخیص جوامع

در این پایان نامه از دو الگوریتم برای گراف‌های وزن‌دار استفاده شده‌است که هر کدام در زیر بخش‌های جداگانه به صورت خلاصه تشریح خواهند شد.

---

[87] Community Structure

[88] Community

۱. انتشار برچسب نامتقارن[۸۹]

۲. الگوریتم نیومن-گردی-کالت[۹۰]

الگوریتم انتشار برچسب نامتقارن

الگوریتم انتشار برچسب، یک الگوریتم محلی برای شناسایی است، که از ساختار محلی و توپولوژیکی خود شبکه کمک می‌گیرد. در الگوریتم انتشار برچسب، در ابتدا هر گره با یک مقدار واحد مقدار دهی اولیه می‌شود و در هر چرخه الگوریتم به هر یک برچسب اختصاص داده می‌شود به گونه‌ای که همسایگان آن گره همان برچسب را دریافت می‌کنند. این روال تکرار می‌شود تا همه گره‌ها برچسب‌دار شوند. در پایان گره‌های با برچسب یکسان ادغام می‌شوند. در الگوریتم انتشار برچسب نامقارن، بروزرسانی گره‌ها ناهمزمان است یعنی هر گره بروزرسانی می‌شود بدون اینکه منتظر بقیه گره‌ها شود [۴۹].

الگوریتم نیومن-گردی-کالت

این الگوریتم، الگوریتم بهینه سازی شده نیومن-گروین هستش که پیچیدگی زمانی کمتری دارد ومی‌تواند شبکه‌های با تعداد گره‌ها و یال‌ها زیاد را پارتیشن‌بندی کند [۵۰]. این الگوریتم برای گراف‌های وزن‌دار و بدون وزن استفاده می‌شود.

پیچیدگی این الگوریتم به صورت زیر است:

$$O(md \log n) \qquad (۱۳-۲)$$

n تعداد یال‌ها، m تعداد گره‌ها و d عمق نمودار دندروگرام است.

۲-۱۱- معیارهای ارزیابی

معیارهای ارزیابی روش پیشنهادی اول

در حوزه هوش مصنوعی، ماتریس درهم ریختگی[۹۱] به ماتریسی گفته می‌شود که در آن عملکرد الگوریتم‌های مربوطه را نشان می‌دهند[۵۱]. معمولاً چنین نمایشی برای الگوریتم‌های یادگیری با ناظر استفاده می‌شود، اگرچه در یادگیری بدون ناظر نیز کاربرد دارد. معمولاً به کاربرد این ماتریس در الگوریتم‌های بدون ناظر ماتریس تطابق می‌گویند. هر ستون از ماتریس، نمونه‌ای از مقدار پیش‌بینی‌شده را نشان می‌دهد. درصورتی‌که هر سطر نمونه‌ای واقعی (درست) را در بر دارد. اسم این ماتریس نیز ازآنجا بدست می‌آید که امکان اشتباه و تداخل بین نتایج را

---

[89] Asynchronous Label Propagation

[90] Clauset-Newman-Moore Greedy Modularity

[91] Confusion Matrix

آسان‌تر می‌توان مشاهده کرد.

جدول ۲—۱ماتریس درهم ریختگی

		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

درنهایت نتایج به‌دست‌آمده مورد ارزیابی قرار گرفته و برای موارد مختلف تفسیر و استفاده می‌شود. در ارزیابی معمولا معیارهای زیر متصور است.

- **دقت**[۹۲]

به طور کلی، دقت به این معناست که مدل تا چه اندازه خروجی را درست پیش‌بینی می کند. با نگاه کردن به دقت، بلافاصله می‌توان دریافت که آیا مدل درست آموزش دیده است یا خیر و کارآیی آن به طور کلی چگونه است. اما این معیار اطلاعات جزئی در مورد کارآیی مدل ارائه نمی‌دهد.

$$acc = \frac{tp + tn}{tp + fp + tn + fn}$$

- **صحت**[۹۳]

وقتی که مدل نتیجه را مثبت[۹۴] پیش‌بینی می کند، این نتیجه تا چه اندازه درست است؟ زمانی که ارزش False Positives بالا باشد، معیار صحت، معیار مناسبی خواهد بود

$$prc = \frac{tp}{tp + fp}$$

- **فراخوانی یا حساسیت** [۹۵]

در نقطه مقابل این پارامتر، ممکن است در مواقعی دقت تشخیص کلاس منفی حائز اهمیت باشد. از متداول‌ترین پارامترها

---

[92] Accuracy

[93] Precision

[94] Positive

[95] Recall

که معمولا در کنار حساسیت بررسی می‌شود، پارامتر خاصیت[96] است که به آن «نرخ پاسخ‌های منفی درست[97]» نیز می گویند. خاصیت به معنی نسبتی از موارد منفی است که آزمایش آن‌ها را به درستی به عنوان نمونه منفی تشخیص داده است. این پارامتر به صورت زیر محاسبه می‌شود.

$$rec = \frac{tp}{tp + fn}$$

- **معیار ارزیابی F1[98]**

معیار F1، یک معیار مناسب برای ارزیابی دقت یک آزمایش است. این، معیار Precision و Recall را با هم در نظر می گیرد. معیار F1 در بهترین حالت، یک و در بدترین حالت صفر است.

$$f1 = \frac{2 \times prc \times rec}{prc + rec}$$

معیارهای ارزیابی روش پیشنهادی دوم

برای ارزیابی عملکرد الگوریتم‌های تشخیص جوامع از سه تابع زیر برای محاسبه کیفیت جوامع استفاده کرده‌ایم. که هر کدام را به صورت خلاصه تشریح خواهیم کرد.

- **تابع کیفیت Modularity**

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(C_i, C_j).$$

(۱۴-۲)

A ماتریس مجاورت، m تعداد کل یال‌ها گراف، $k_ik_j=2m$ تعداد یال‌های بین گره i و گره j است، اگر$C_j=C_i$باشد تابع δ یک می‌شود در غیر این صورت صفر می‌شود [۵۲].

- **تابع کیفیت Performance**

تابع کیفیت کارایی یا پرفورمنس یکی از بهترین توابع ارزیابی پارتیشن‌های درست شده توسط الگوریتم‌ها است. خروجی این الگوریتم باید بین ۱ > ۰< ۱- باشد. میزان نزدیکی خروجی به عدد به معنی تشخیص پارتیشن خوب

---

[96] Specificity

[97] True Negative Rate

[98] F1-Score

توسط الگوریتم است [۵۲]. i و j شماره گره‌ها هستند، E مجموعه یال‌ها، C جامعه یا پارتیشن، n تعداد کل گره‌ها است.

$$P(\mathcal{P}) = \frac{|\{(i,j) \in E, C_i = C_j\}| + |\{(i,j) \notin E, C_i \neq C_j\}|}{n(n-1)/2}. \quad (۱۵-۲)$$

- **تابع کیفیت Coverage**

تعداد یال‌های داخلی جامعه نسبت به تعداد کل یال‌های گراف، خروجی باید بین صفر تا یک باشد، نزدیکی بیشتر این کمیت به مقدار یک به معنی تعداد یال‌های بیشتر درون جامعه است [۵۲].

$$P(P) = \text{Intra-Community-Edges} / \text{Total-Edges} \quad (۱۶-۲)$$

۲-۱۲- جمع‌بندی

در این فصل ابتدا به تفصیل هنجاری، تشخیص ناهنجاری، کاربردها و روش‌های آن را تشریح کردیم و در ادامه در مورد پیشگویی پیوند، انواع پیشگویی پیوند، کاربردها و روش‌های آن توضیح دادیم و همچنین تشیخص ناهنجاری یال در گراف و روش‌های که برای تشخیص ناهنجاری یال در گراف استفاده شده بود را بیان کردیم و در بخش آخر الگوریتم تجزیه پیازی،لاین گراف، تشخیص جوامع، توابع کیفیت یا توابع ارزیابی جوامع که در این پایان نامه استفاده شده است را تشریح کردیم.

فصل سوم: روش پیشنهادی

۳-۱- مقدمه

در این فصل تحقیقات انجام شده در زمینه تشخیص ناهنجاری یال در قالب دو روش پیشنهادی به صورت بدون نظارت ارائه می‌شود. روش پیشنهادی اول برای گراف‌های بدون وزن و روش دوم برای گراف‌های وزن‌دار ارائه شده است. در روش پیشنهادی اول، ابتدا یال‌های ناهنجار(دیتای نویز) به گراف اضافه می‌کنیم سپس پیشگویی پیوند منفی بر اساس الگوریتم‌های جاکارد، آدامیک-آدار، پیوست امتیازی و همسایه‌های مشترک روی گراف اعمال و یال‌های ناهنجار گراف تشخیص داده می‌شود. در روش پیشنهادی دوم الگوریتم‌های پیشگویی پیوند منفی روی گراف اعمال و ضعیف‌ترین یال‌ها از گراف حذف می‌شود، بعد از حذف یال‌های ناهنجار، الگوریتم تشخیص جوامع روی گراف اعمال شده و نتایج ارزیابی می‌شود. روش جدیدی برای پیشگویی پیوند منفی گراف‌های وزن‌دار ارایه شده است و همچنین روش جدیدی هم برای تولید دیتای نویز پیشنهاد گردیده است.

۳-۲- روش پیشنهادی اول (حذف ناهنجاری با پیشگویی پیوند منفی در گراف بدون وزن)

در این روش، برای حذف یال‌های ناهنجار در گراف‌های بدون وزن از روش پیشگویی پیوند منفی استفاده شده است. روش پیشگویی پیوند منفی، برای گراف‌های بدون وزن به صورتی است که ماتریس مجاورت نات[۹۹](یعنی صفرها به یک و یک‌ها به صفر تبدیل می‌شوند) شده و چهار الگوریتم پیشگویی پیوند بدون نظارت (AA, PA, JC, CN) روی آن اعمال می‌شود، قویترین یالی که پیشگویی شده است در واقع ضعیف‌ترین یال گراف است، که به عنوان ناهنجاری در نظر گرفته می‌شود.

برای ارزیابی این روش ما با استفاده از یک روش که در فصل چهارم تشریح شده یال‌های ناهنجار تولید(دیتای نویز) و به گراف اضافه می‌کنیم. بعد پیشگویی پیوند منفی را روی مجموعه داده اعمال و نتایج را با چهار معیار ارزیابی معروف ارزیابی می‌کنیم. فلوچارت (شبه الگوریتم) روش پیشنهادی اول (حذف ناهنجاری با پیشگویی پیوند منفی در گراف بدون وزن) در شکل ۳-۱ دیده می شود:

[99] Not

Input:

G: Plain graph dataset

Output:

G: Plain graph without anomaly edges

Begin

1. // Socre Function{PA, AA,CN,JC}
2. G = Reading dataset
3. M = Convert graph to matrix
4. IM = Not matrix M
5. LP = Apply Socre Function on M
6. SLP = Descending sort list LP
7. R = Remove Anomaly edges SLP from G
8. Return G

End

شکل ۳—۱ شبه کد روش پیشنهادی اول برای تشخیص ناهنجاری یال در گراف‌های بدون وزن

نمودار شبه الگوریتم روش پیشنهادی اول در شکل ۳-۲ نمایش داده شده است.

شکل ۳—۲ نمودار شماتیک روش پیشنهادی اول

۳-۳- روش پیشنهادی دوم (حذف ناهنجاری با پیشگویی پیوند منفی در گراف وزن‌دار)

گراف وزن‌دار، گرافی است که در آن به هر یال، یک عدد (وزن) اختصاص داده می‌شود. این وزن‌ها بسته به مشکل موجود یا نوع مسئله ممکن است نمایانگر هزینه، مسافت، طول یا ظرفیت باشد. چنین گراف‌هایی در زمینه‌های مختلف مانند، پیدا کردن کوتاهترین مسیر در مسئله فروشنده دوره گرد کاربرد دارد. گراف‌های وزن‌دار در دنیای واقعی بسیار پرکاربرد هستند. از این رو اهمیت ویژه‌ای در بین پژوهشگران دارد. در شبکه‌های بدون وزن با نات کردن ماتریس مجاورت(ماتریس شبکه‌های بدون وزن بولی هستند) می‌توان پیشگویی پیوند منفی انجام داد. اما این روش برای گراف‌های وزن‌دار غیر قابل استفاده است چون ماتریس اوزان گراف، بولی (صفر و یک) نیست تا آن را معکوس کنیم، ماتریس گراف‌های وزن‌دار ماتریسی از اوزان است. برای همین ما روشی ساده ولی کاربردی برای پیشگویی پیوند منفی در گراف‌های وزن‌دار ارائه داده‌ایم که در ادامه به تشریح آن می‌پردازیم. برای اعمال پیشگویی پیوند منفی روی گراف‌های وزن‌دار باید وزن یال‌ها را معکوس کنیم، یعنی، قوی‌ترین یال، به ضعیف‌ترین یال و ضعیف‌ترین یال، به قوی‌ترین یال تبدیل شود. ما این کار را به صورت زیر انجام داده‌ایم: وزن یالی که دارای کوچک‌ترین وزن در گراف است را با وزن یالی که بیشترین وزن را در گراف دارد جمع می‌شود و حاصل جمع آن‌ها از همه یال‌ها کم می‌شود، در این صورت قوی‌ترین یال به ضعیف‌ترین یال و ضعیف‌ترین یال به قوی‌ترین یال تبدیل می‌شود (فرمول ۳-۱). اگر با این روش وزن یال‌های موجود را معکوس کنیم و الگوریتم‌های پیشگویی پیوند را روی آن اعمال کنیم، قوی‌ترین یالی که پیشگویی شده است به معنی ضعیف‌ترین یال ناهنجار است که باید حذف شود.

$$E(v,u) = \big(maxWeigh(G) + minWeight(G)\big) - weight(E(u,v)) \qquad (۱-۳)$$

شبه الگوریتم روش پیشنهادی دوم (حذف ناهنجاری با پیشگویی پیوند منفی در گراف وزن‌دار):

Input:

G: Weighted graph dataset

P: Percentage remove anomaly edge form graph

Output:

G: Weighted graph without anomaly edge

Begin

1. // Socre Function{PA, AA,CN,JC}
2. G = Reading dataset
3. NumEdge = Calculate percentage graph [P, G]
4. M = Convert graph to matrix
5. IM = Inverse weighted edge matrix with Feq (1-3)

6. LP = Apply Socre Function on M
7. SLP = Descending sort list LP
8. For i=0 to NumEdge
9. Remove edge SLP[i] form G
10. Return G

End

شکل ۳—۳شبه کد روش پیشنهادی دوم برای تشخیص ناهنجاری یال در گراف‌های وزن‌دار

در شکل ۳-۴ نمودار شماتیک روش پیشنهادی دوم نشان داده شده است.

شکل ۳—۴ نمودار شماتیک، الگوریتم پیشگویی پیوند منفی برای تشخیص ناهنجاری یال در گراف‌های وزن‌دار

۳-۴- جمع‌بندی

در این فصل دو روش برای تشخیص ناهنجاری یال ارائه گردید. در روش اول با استفاده از پیشگویی پیوند منفی یال‌های ناهنجاری تشخیص داده شد.در روش دوم، روشی جدید برای پیشگویی پیوند منفی در گراف‌های وزن‌دار ارائه شد و با استفاده از آن یال‌های ناهنجار از گراف حذف و باعث بهبود جوامع می‌شود. در فصل بعدی عملکرد هر دو روش را روی مجموعه داده‌ها مختلف بررسی و ارزیابی می کنیم.

فصل چهارم: نتایج و تفسیر

۴-۱- مقدمه

هدف از این فصل ارزیابی عملکرد روش‌های پیشنهادی است. روش پیشنهادی اول برای گراف‌های بدون وزن و روش پیشنهادی دوم برای گراف‌های وزن‌دار ارائه شده است. در هر دو روش پیشنهادی ، از چهار الگوریتم پیشگویی پیوند منفی(AA, CN, PA, JA) به صورت مجزا برای تشخیص یال‌های ناهنجار استفاده شده است. نتایج روش پیشنهادی اول با هشت الگوریتم دیگر مقایسه شده است. از این هشت الگوریتم، چهار الگوریتم برای تشخیص ناهنجاری یال استفاده ارائه شده و چهار الگوریتم دیگر برای تشخیص گره ارائه شده که ما با استفاده از لاین گراف، یال‌های ناهنجار را با این چهار الگوریتم تشخیص گره حذف می‌کنیم. روش مقایسه برای روش اول به صورت زیر است: قبل از هر چیز ما به گراف، یال‌های ناهنجار (دیتای نویز) اضافه کرده و بعد روش پیشنهادی و سایر الگوریتم‌ها را به صورت مجزا روی گراف اجرا می‌کنیم و نتایج را با چهار معیار ارزیابی معروف یعنی دقت، صحت، فراخوانی و معیار F1، می‌سنجیم. در بخش دوم آزمایش روش پیشنهادی اول، جهت ارزیابی کارایی الگوریتم‌ها به صورت زیر عمل کرده‌ایم: به گراف سه درصد یال ناهنجار اضافه کرده و الگوریتم‌ها را روی آن اعمال و نتایج را ثبت و دوباره سه درصد دیگر اضافه کرده‌ایم و این روال را تکرار شده تا بیست‌ویک درصد یال ناهنجار اضافه و نتایج ثبت شده است و دقت الگوریتم‌ها را با نمودار خطی نشان داده‌ایم. روش پیشنهادی دوم را که برای گراف‌های وزن‌دار ارائه شده است به صورت زیر ارزیابی کرده‌ایم: بر روی هر مجموعه داده با استفاده از الگوریتم‌های پیشگویی پیوند منفی، درصدی از یال‌های ناهنجار (این درصد نسبت به مجموعه داده‌های مختلف متغیر است) را حذف کرده و نتایج را ثبت و آن را با توابع کیفیت جوامع، ارزیابی کرده‌ایم.

هر روش پیشنهادی روی چهار مجموعه داده اجرا شده است که نتایج به صورت نمودار گزارش شده است. در آخر هر روش پیشنهادی برترین الگوریتم‌ها برای هر مجموعه داده مشخص شده است و توضیحاتی در مورد آن‌ها ارائه شده است.

۴-۱-۱- نتایج علمی

در این قسمت، نتایج آزمایشات هر دو روش پیشنهادی ارائه می‌گردد و از مجموعه داده‌های زیر استفاده شده است.

مجموعه داده‌های بدون وزن استفاده شده برای روش پیشنهادی اول:

جدول ۴—۱ مجموعه داده‌های استفاده شده برای گراف‌های بدون وزن

نام مجموعه داده	تعداد گره‌ها	تعداد یال‌ها	چگالی
Dolphins	۶۲	۱۵۹	۰,۰۸۴
Jazz	۱۱۵	۶۱۳	۰,۰۹۳
Email	۱۱۳۴	۱۱۳۳	۰,۰۰۱
Trinity100	۲۶۱۳	۱۱۱۹۹۶	۰,۰۳۲

مجموعه داده‌های وزن‌دار استفاده شده برای روش پیشنهادی سوم و چهارم:

جدول ۴—۲ مجموعه داده‌های استفاده شده برای گراف‌های وزن‌دار

نام مجموعه داده	تعداد گره‌ها	تعداد یال‌ها	چگالی
Lesmis[100]	۷۷	۲۵۴	۰,۰۸۶
King James[101]	۱۷۷۳	۹۱۳۱	۰,۰۰۵
Netscience[102]	۱۴۶۱	۲۷۴۲	۰,۰۰۲
Adolescent[103]	۲۵۳۹	۱۰۴۵۵	۰,۰۰۳

۴-۲- روش تولید یال ناهنجار(دیتای نویز)

برای آزمایش و مقایسه روش پیشنهادی با سایر الگوریتم‌ها در گراف‌های بدون وزن، نیاز داشتیم که با استفاده از یک روش، دیتای نویز به گراف اضافه کنیم و روش پیشنهادی و سایر الگوریتم‌ها را روی آن اجرا کنیم و با معیارهای ارزیابی، کارایی این الگوریتم‌ها را بسنجیم و مقایسه کنیم. چون نوع تشخیص ناهنجاری که ما بر روی آن کار می کنیم مبتنی بر ساختار است، برای تولید دیتای نویز به روش زیر عمل کردیم: الگوریتم تجزیه پیازی را روی گراف اعمال شده و ناهنجارترین گره‌ها را مشخص و چند درصد یال ناهنجار(بسته به چگالی گراف این درصد متغییر است) به گراف اضافه می‌شود. بعد دوباره الگوریتم تجزیه پیازی اعمال و بازم اضافه کردن چند درصد یال به گره‌های ناهنجار اضافه می‌شود، این روال ادامه پیدا می کند تا ۱۰ درصد یال ناهنجار(۱۰ درصد نسبت به چگالی گراف) به گراف اضافه می‌شود. بعد الگوریتم‌ها روی گراف اعمال و نتایج با معیارهای ارزیابی دقت، صحت، فراخوانی و معیار اف ۱ ارزیابی شده است.

---

[100] http://networkrepository.com/lesmis.php

[101] http://www.linkprediction.org/index.php/link/resource/data

[102] http://www.linkprediction.org/index.php/link/resource/data

[103] http://www.linkprediction.org/index.php/link/resource/data

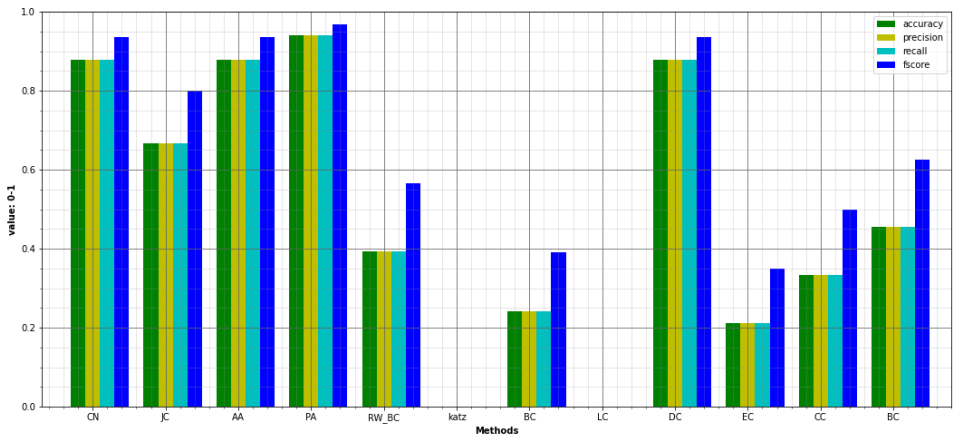
این روش تولید دیتای نویز فقط برای گراف‌های بدون وزن طراحی شده است و برای گراف‌های وزن‌دار جواب نمی‌دهد. در گراف‌های وزن‌دار علاوه بر ساختار گراف، وزن یال‌ها هم در ناهنجار بدون گره یا یال تاثیر گذار هستند. برای همین از روش دیگه‌ای برای ارزیابی و عملکرد روش‌پیشنهادی دوم(گراف‌های وزن دار) استفاده کرده‌ایم.

۴-۳- نتایج روش پیشنهادی اول (حذف ناهنجاری با پیشگویی پیوند منفی در گراف بدون وزن)

در این بخش نتایج روش پیشنهادی اول را که بر روی چهار مجموعه داده جدول ۴-۲ اجرا کرده ایم، گزارش می‌شود. در این روش نتایج به سه بخش تقسیم بندی می‌شود: در قسمت اول ۱۰درصد یال ناهنجار به هر مجموعه داده اضافه شده است و با معیارهای ارزیابی کارایی روش پیشنهادی و سایر الگوریتم‌ها به صورت نمودار میله‌ای مقایسه شده است. درقسمت دوم نتایج در یک جدول گردآوری شده و مقایسه شده است. در قسمت سوم هر بار ۳ درصد یال ناهنجار به گراف اضافه شده است و معیار ارزیابی دقت برای همه الگوریتم‌ها محاسبه و بعد ۳ درصد دیگه اضافه و معیار دقت محاسبه شده است، این روال ادامه پیدا می کند تا اینکه ۲۱ درصد(نسبت به چگالی مجموعه داده) یال اضافه می‌شود، نتایج این قسمت به صورت نمودار خطی ارائه شده است.

در جدول اختصارات، نام کامل الگوریتم‌ها ثبت شده است. در همه نمودارهاCN, JC, AA, PA روش‌های پیشنهادی هستند. الگوریتم‌های RW_BC, Katz, BC, LC روش‌های تشخیص ناهنجاری یال هستند. روش‌های DC, EC, CC, BC روش‌های تشخیص ناهنجاری گره هستند که با استفاده از لاین گراف، گراف را برای این الگوریتم‌ها برعکس کردیم یعنی یال را به گره و گره را به یال تبدیل کرده‌ایم تا بتوانند یال‌های ناهنجار را حذف کنند.

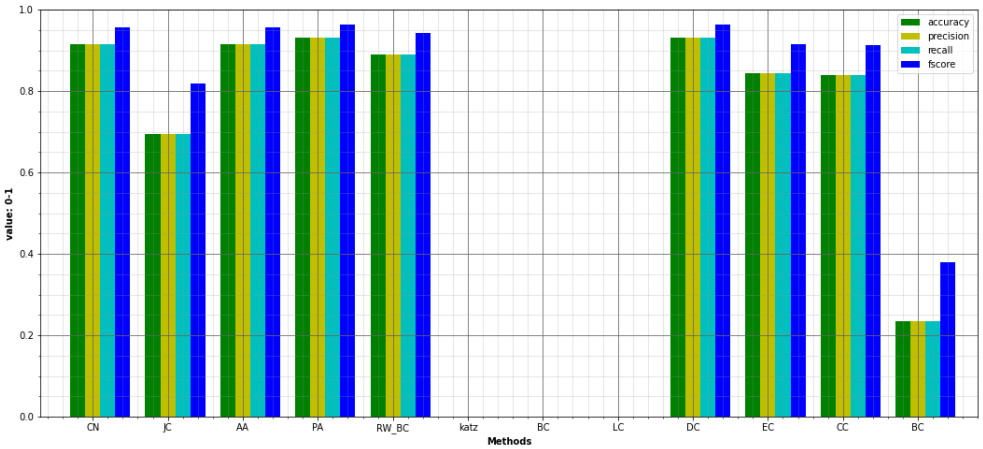
۴-۳-۱- نتایج مجموعه داده Dolphins



شکل ۴—۱ مقایسه نتایج روش پیشنهادی اول و الگوریتم‌های رقیب روی مجموعه داده Dolphins

در مجموعه داده دلفین، PA بهتر از بقیه بوده و بعد آن DC بهتر عمل کرده است. همچنین ktz و LC ضعیف‌ترین عملکرد را داشته است. در کل روش پیشنهادی در یافتن یال‌های ناهنجار عملکرد خوبی نسبت به سایری الگوریتم‌ها داشته است.

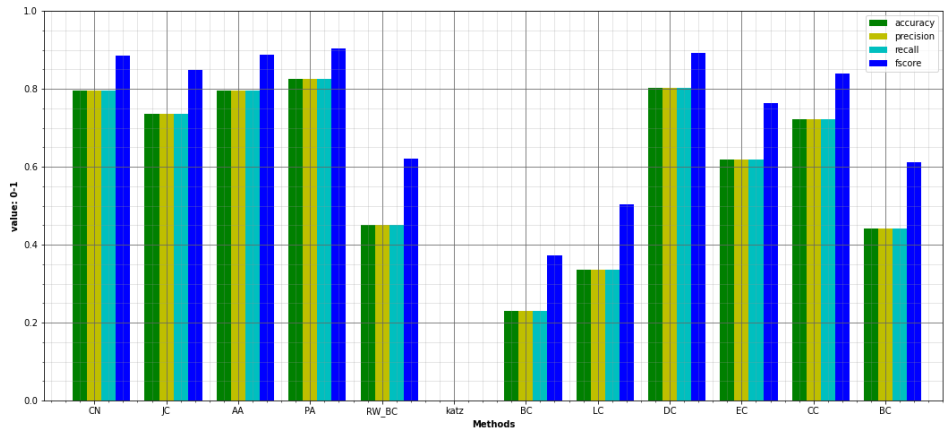
۴-۳-۲- نتایج مجموعه داده Jazz



شکل ۴—۲ مقایسه نتایج روش پیشنهادی اول و الگوریتم‌های رقیب روی مجموعه داده Jazz

در مجموعه داده جاز PA و DC عملکردی نزدیک بهم داشته‌اند یعنی در یافتن یال‌های ناهنجار با دقت بالای ۹۵ درصد توانستند نسبت به سایر الگوریتم‌ها بهتر باشند. الگوریتم CN، AA و RW_BC هم با دقت بالای ۹۰ درصد یال‌های ناهنجار را پیدا کرده‌اند. همان طور که از نمودار مشخص است روش‌های ktz، BC و LC ضعیف‌ترین عملکرد را داشته‌اند.

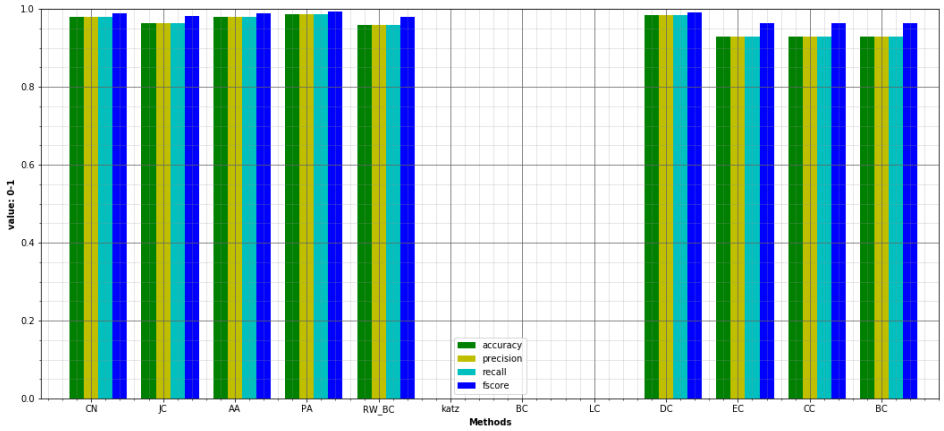
۴-۳-۳- نتایج مجموعه داده Email



شکل ۴—۳ مقایسه نتایج روش پیشنهادی اول و الگوریتم‌های رقیب روی مجموعه داده Email

در مجموعه داده ایمیل روش PA بهترین عملکرد را داشته است بعد الگوریتم‌های CN، AA و DC نسبت به سایر الگوریتم‌ها عملکرد بهتری داشته است. همچنین در این مجموعه داده ضعیف‌ترین روش ktz بوده است.

۴-۳-۴- نتایج مجموعه داده Trinity100



شکل ۴—۴ مقایسه نتایج روش پیشنهادی اول و الگوریتم‌های رقیب روی مجموعه داده *Trinity100*

مجموعه داده Trinity100، یکی از مجموعه داده‌های فیسبوک می‌باشد. در این مجموعه داده الگوریتمPA بهترین نتیجه را داشته است و بعد آن الگوریتم DC بهترین نتیجه را داشته و بعد آن الگوریتم CN بهترین بوده است. الگوریتم‌های Ktz، BC و LC بدترین نتیجه را داشته است.

جدول ۴-۳ مقایسه روش پیشنهادی با سایر الگوریتم‌ها است که شامل پنج ستون اصلی است که چهار ستون آن مجموعه داده‌ها و ستون آخر میانگین نتیجه چهار مجموعه داده برای هر کدام از روش‌ها است و هر ستون اصلی شامل چهار زیر ستون می‌شود که هر ستون آن نشان دهنده یکی از معیارهای ارزیابی دقت (ACC)، صحت (PRE)، فراخوانی یا حساسیت (REC) و معیار F است. قسمتی که بولد شده و زیر آن خط کشیده شده، نشان دهنده قوی‌ترین الگوریتم است. قوی‌ترین الگوریتم PA (روش پیشنهادی) بوده است و بعد آن الگوریتم DC بهترین بوده است.

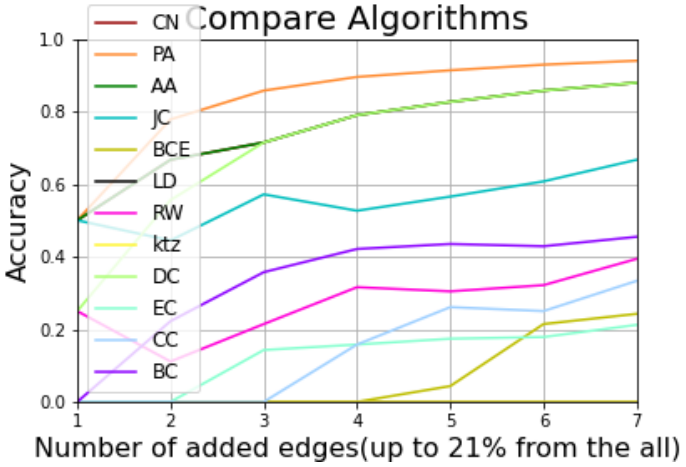
	Dolphins				Jazz				Email				Trinity100				average			
	ACC	PRE	REC	F	ACC	PRE	REC	F	ACC	PRE	REC	F	ACC	PRE	REC	F	ACC	PRE	REC	F
CN	۰.۷۳۳	۰.۷۳۳	۰.۷۳۳	۰.۸۴۶	۰.۹۱۶	۰.۹۱۶	۰.۹۱۶	۰.۹۵۶	۰.۷۹۴	۰.۷۹۴	۰.۷۹۴	۰.۸۸۵	۰.۹۷۸	۰.۹۷۸	۰.۹۷۸	۰.۹۸۹	۰.۸۵۵	۰.۸۵۵	۰.۸۵۵	۰.۹۱۹
JA	۰.۵۳۳	۰.۵۳۳	۰.۵۳۳	۰.۵۳۳	۰.۶۹۳	۰.۶۹۳	۰.۶۹۳	۰.۸۱۸	۰.۸۸۵	۰.۸۸۵	۰.۸۸۵	۰.۸۴۷	۰.۹۶۲	۰.۹۶۲	۰.۹۶۲	۰.۹۸۰	۰.۷۸۵	۰.۷۸۵	۰.۷۸۵	۰.۸۱۹۴
AA	۰.۷۳۳	۰.۷۳۳	۰.۷۳۳	۰.۸۴۶	۰.۹۱۶	۰.۹۱۶	۰.۹۱۶	۰.۹۵۶	۰.۷۹۶	۰.۷۹۶	۰.۷۹۶	۰.۸۸۶	۰.۹۷۸	۰.۹۷۸	۰.۹۷۸	۰.۹۸۹	۰.۸۵۵	۰.۸۵۵	۰.۸۵۵	۰.۹۱۹۲
PA	**۰.۸٤٦**	**۰.۸٤٦**	**۰.۸٤٦**	**۰.۹۲۸**	**۰.۹۳۰**	**۰.۹۳۰**	**۰.۹۳۰**	**۰.۹٦٤**	**۰.۸۸٦**	**۰.۸۸٦**	**۰.۸۸٦**	**۰.۹۰٤**	**۰.۹۸۵**	**۰.۹۸۵**	**۰.۹۸۵**	**۰.۹۹۲**	**۰.۹۱۱**	**۰.۹۱۱**	**۰.۹۱۱**	**۰.۹۲۰**
RW	۰.۲۶۶	۰.۲۶۶	۰.۲۶۶	۰.۴۲۱	۰.۸۹۰	۰.۸۹۰	۰.۸۹۰	۰.۹۴۲	۰.۴۵۱	۰.۴۵۱	۰.۴۵۱	۰.۶۲۱	۰.۹۵۹	۰.۹۵۹	۰.۹۵۹	۰.۹۷۹	۰.۶۴۱	۰.۶۴۱	۰.۶۴۱	۰.۶۹۸
LC	۰.۰	۰.۰	۰.۰	۰.۰	۰.۰	۰.۰	۰.۰	۰.۰	۰.۳۳۵	۰.۳۳۵	۰.۳۳۵	۰.۵۰۲	۰.۰	۰.۰	۰.۰	۰.۰	۰.۳۳۵	۰.۳۳۵	۰.۳۳۵	۰.۵۰۲
BC	۰.۰	۰.۰	۰.۰	۰.۰	۰.۰	۰.۰	۰.۰	۰.۰	۰.۲۲۹	۰.۲۲۹	۰.۲۲۹	۰.۳۷۳	۰.۰	۰.۰	۰.۰	۰.۰	۰.۲۲۹	۰.۲۲۹	۰.۲۲۹	۰.۳۷۳
Ktz	۰.۰	۰.۰	۰.۰	۰.۰	۰.۰	۰.۰	۰.۰	۰.۰	۰.۱۷	۰.۱۷	۰.۱۷	۰.۳۴	۰.۰	۰.۰	۰.۰	۰.۰	۰.۱۷	۰.۱۷	۰.۱۷	۰.۳۴
DC	۰.۷۳۳	۰.۷۳۳	۰.۷۳۳	۰.۸۴۶	۰.۹۳۰	۰.۹۳۰	۰.۹۳۰	۰.۹۶۴	۰.۸۰۳	۰.۸۰۳	۰.۸۰۳	۰.۸۰۳	۰.۹۸۴	۰.۹۸۴	۰.۹۸۴	۰.۹۹۱	۰.۸۶۲	۰.۸۶۲	۰.۸۶۲	۰.۹۰۱
EC	۰.۱۳۳	۰.۱۳۳	۰.۱۳۳	۰.۱۳۳	۰.۸۴۳	۰.۸۴۳	۰.۸۴۳	۰.۹۱۴	۰.۶۱۸	۰.۶۱۸	۰.۶۱۸	۰.۷۶۴	۰.۹۲۷	۰.۹۲۷	۰.۹۲۷	۰.۹۶۲	۰.۶۳۰	۰.۶۳۰	۰.۶۳۰	۰.۶۹۳
CC	۰.۰	۰.۰	۰.۰	۰.۰	۰.۸۳۹	۰.۸۳۹	۰.۸۳۹	۰.۹۱۲	۰.۷۲۲	۰.۷۲۲	۰.۷۲۲	۰.۸۳۹	۰.۹۲۰	۰.۹۲۰	۰.۹۲۰	۰.۹۵۹	۰.۶۲۰	۰.۶۲۰	۰.۶۲۰	۰.۶۷۷
BC	۰.۴	۰.۴	۰.۴	۰.۵۷۱	۰.۲۳۳	۰.۲۳۳	۰.۲۳۳	۰.۳۷۸	۰.۴۴۰	۰.۴۴۰	۰.۴۴۰	۰.۶۱۱	۰.۹۱۹	۰.۹۱۹	۰.۹۱۹	۰.۹۵۸	۰.۴۹۸	۰.۴۹۸	۰.۴۹۸	۰.۶۲۹

جدول ۴—۳مقایسه روش پیشنهادی با سایر الگوریتم‌ها

مقایسه دقت الگوریتم‌ها

در این بخش دقت روش پیشنهادی(CN, PA, AA, JC) با هشت الگوریتم دیگر مقایسه شده است. روش مقایسه به این صورت بوده است: سه درصد یال ناهنجار اضافه شده بعد الگوریتم‌ها اعمال و نتایج ثبت شده است بعد سه درصد دیگر یال به گراف اضافه شده و نتایج ثبت اعمال شده است همین روال تکرار شده تا بیست و یک درصد یال به گراف اضافه شده است.

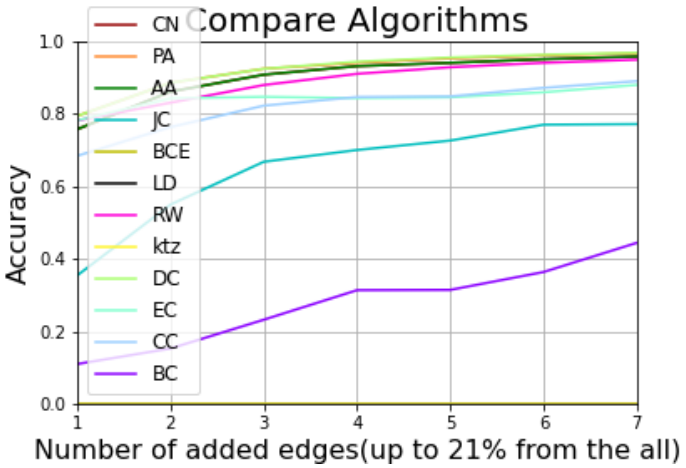
- **مجموعه داده Dolphins**



شکل ۴—۵ مقایسه دقت روش پیشنهادی و الگوریتم‌های رقیب روی مجموعه داده Dolphins

روش PA، بهترین عملکرد را دارد و روش ktz بدترین عملکرد را داشته است.

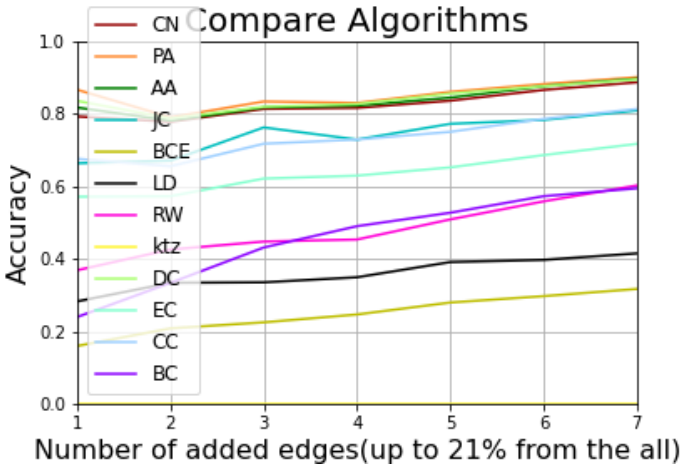
- **مجموعه داده Jazz**



شکل ۴—۶ مقایسه دقت روش پیشنهادی و الگوریتم‌های رقیب روی مجموعه داده Jazz

روش DC و PA بهترین عملکرد را داشته است.

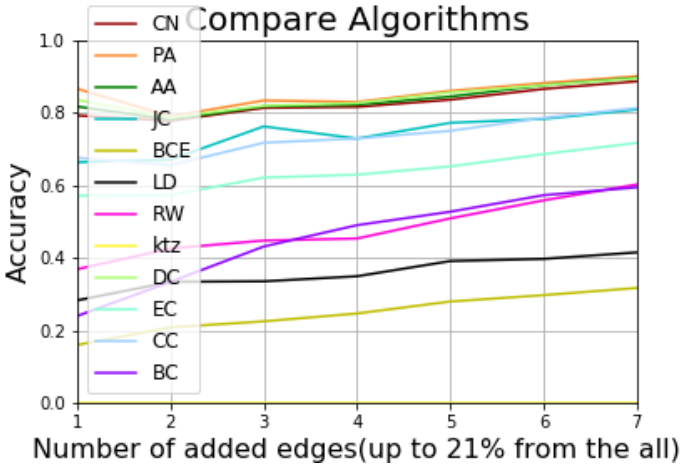
- **مجموعه داده Email**



شکل ۴—۷ مقایسه دقت روش پیشنهادی و الگوریتم‌های رقیب روی مجموعه داده Email

روش PA بهترین عملکرد را داشته است

- **مجموعه داده Trinity100**



شکل ۴—۸ مقایسه دقت روش پیشنهادی و الگوریتم‌های رقیب روی مجموعه داده Trinity100

در این مجموعه داده روش PA، DC و CN خیلی نزدیک به هم بوده‌اند و بهتر از بقیه الگوریتم‌ها بوده‌اند. و ضعیف‌ترین الگوریتم‌ها Ktz و BCE بوده است.

۴-۴- جمع‌بندی و تفسیر روش اول

- بهترین روش PA بوده که در همه مجموعه داده‌ها بهترین عملکرد را داشته است و بعد آن CN بوده است و بعد آن DC
- ضعیف‌ترین روش Ktz بوده است
- روش BC، LC و Ktz برای مجموعه داده‌های کوچک تقریبا کارایی صفر داشته‌اند
- در مجموعه داده کم چگالی(ایمیل) قابلیت پیشگویی کم‌تر بوده است
- روش پیشنهادی ما در مجموعه داده‌های بزرگ‌تر عملکرد بهتری داشته است

۴-۵- نتایج روش پیشنهادی دوم (حذف ناهنجاری با پیشگویی پیوند منفی در گراف وزن‌دار)

در این بخش نتایج روش پیشنهادی دوم را که بر روی ۴ مجموعه داده جدول ۴-۳ اجرا کرده ایم، گزارش می‌شود.

۴-۵-۱- نتایج مجموعه داده Lesmis

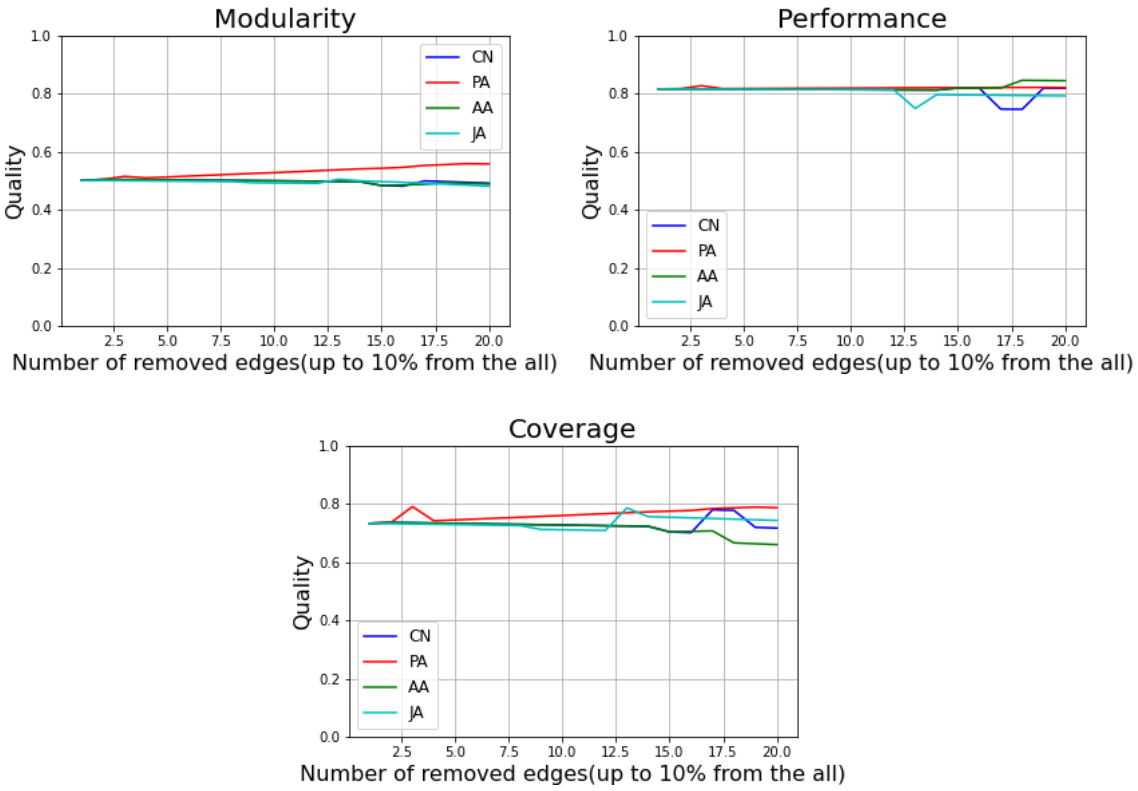
۴-۵-۱-۱- نتایج الگوریتم ALC

شکل ۴—۹ نتایج الگوریتم ALC برای مجموعه داده Lasmis

بر اساس هر سه معیار، PA بهتر از سایر روش‌ها باعث بهبود جوامع شده هر چند نوساناتی داشته است. الگوریتم CN و AA تقریبا

شبیه هم جوامع رو بهبود داده‌اند. بر اساس معیار ماژولاریتی الگوریتم AA ضعیف‌تر از بقیه الگوریتم ها عمل کرده است. در کل می‌توان نتیجه گرفت که روش پیشنهادی تا حدودی باعث بهبود جوامع مجموعه داده Lasmis شده است.

۴-۵-۲-۱- نتایج الگوریتم GMC

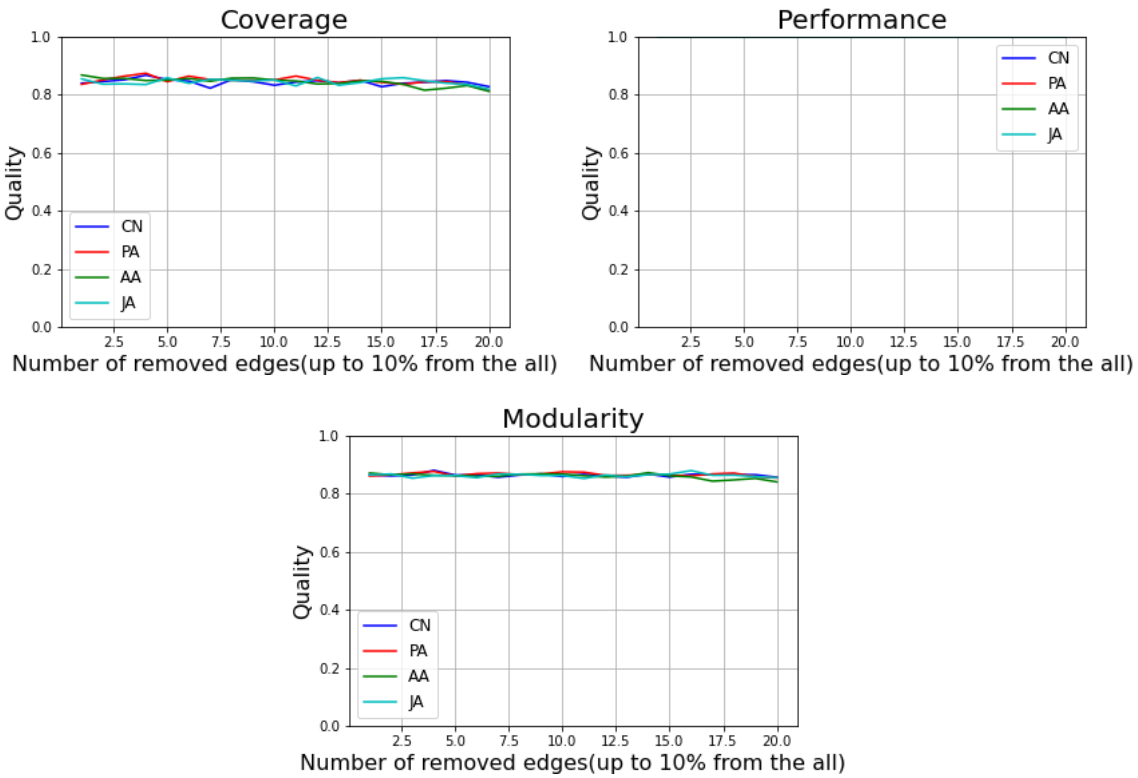


شکل ۴—۱۰ نتایج الگوریتم GMC برای مجموعه داده Lasmis

بر اساس معیار ماژولاریتی و کاوریج، الگوریتم PA بهتر از همه باعث بهبود جوامع شده است. بر اساس معیار کارایی یا پرفورمنس الگوریتم AA بهتر از همه بوده است. اما بر اساس معیار کاوریج الگوریتم AA ضعیف‌تر از همه بوده است. بر اساس معیار ماژولاریتی و کاوریج، روش پیشنهادی باعث بهبود چشم‌گیر جوامع مجموعه داده شده است.

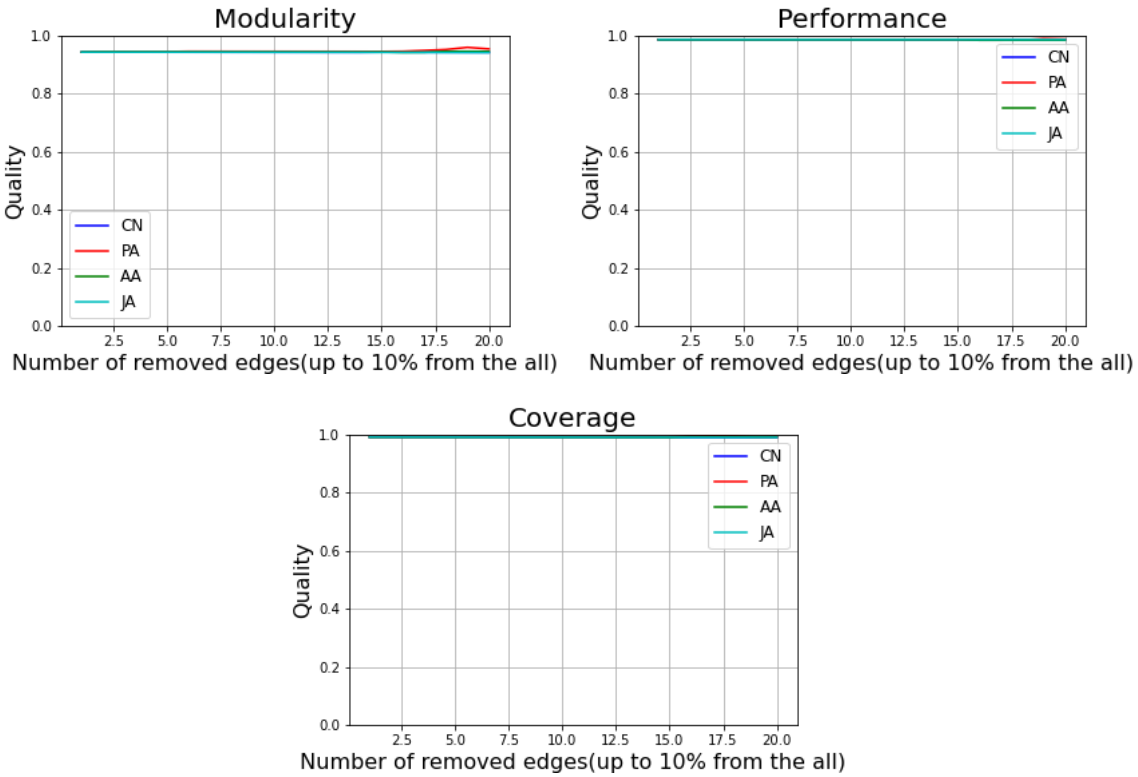
۴-۵-۲- نتایج مجموعه داده Netscience

۴-۵-۲-۱- الگوریتم ALC



شکل ۴—۱۱ نتایج الگوریتم ALC برای مجموعه داده Netscience

۴-۵-۲-۲- نتایج الگوریتم GMC

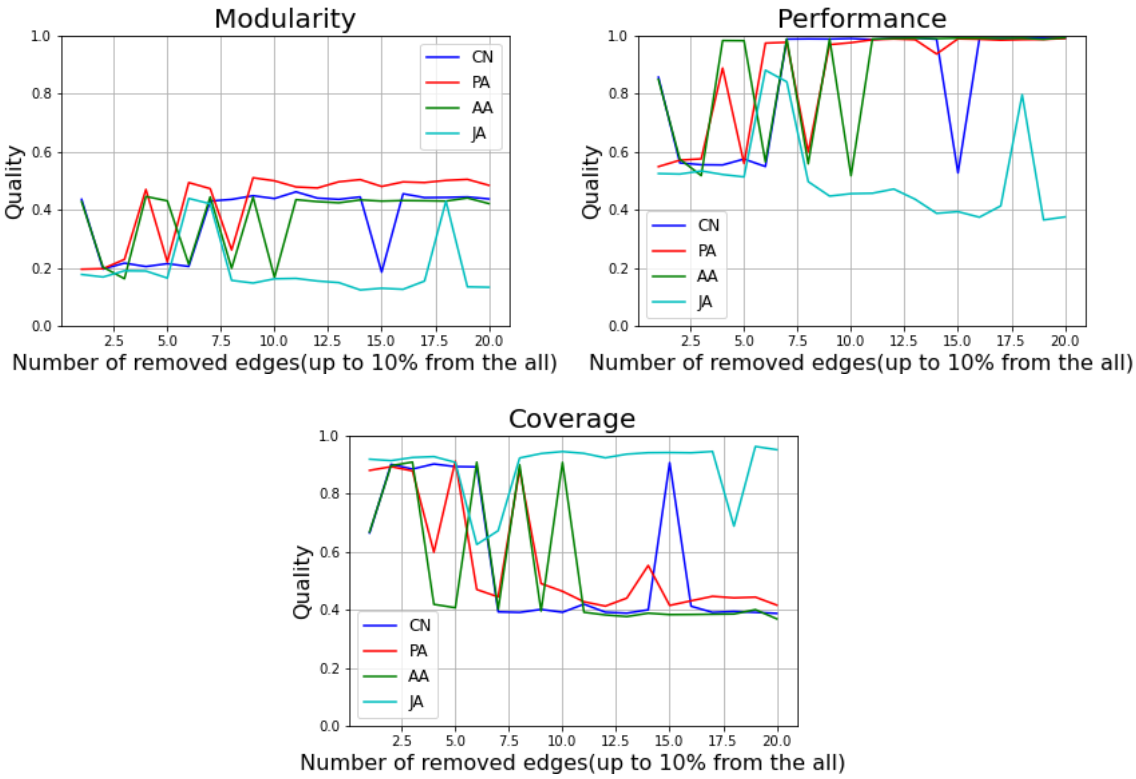


شکل ۴—۱۲ نتایج الگوریتم GMC برای مجموعه داده Netscience

در مجموعه داده Netscience چون جوامع در بالاترین سطح قرار داشته‌اند روش پیشنهادی خیلی کم توانسته جوامع را بهبود بدهد. بر اساس معیار ماژولاریتی روش PA توانسته جوامع را تا حدود بهبود دهد. کار کردن روی همچین مجموعه دادهی برای بهبود دادن خیلی سخت است چون جوامع این مجموعه داده خیلی متراکم است و از کیفیت بالایی برخوردار است.

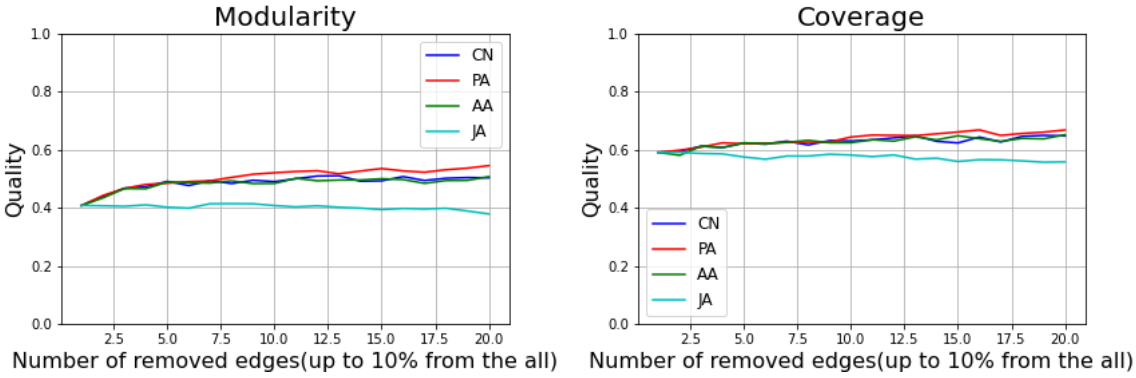
۴-۵-۳- نتایج مجموعه داده King James

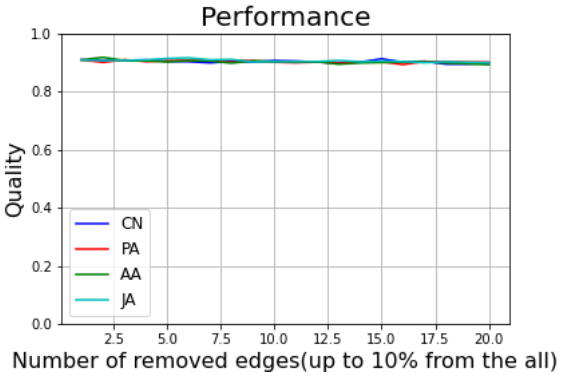
۴-۵-۳-۱- نتایج الگوریتم ALC



شکل ۴—۱۳ نتایج الگوریتم ALC برای مجموعه داده King James

۴-۵-۳-۲- نتایج الگوریتم GMC



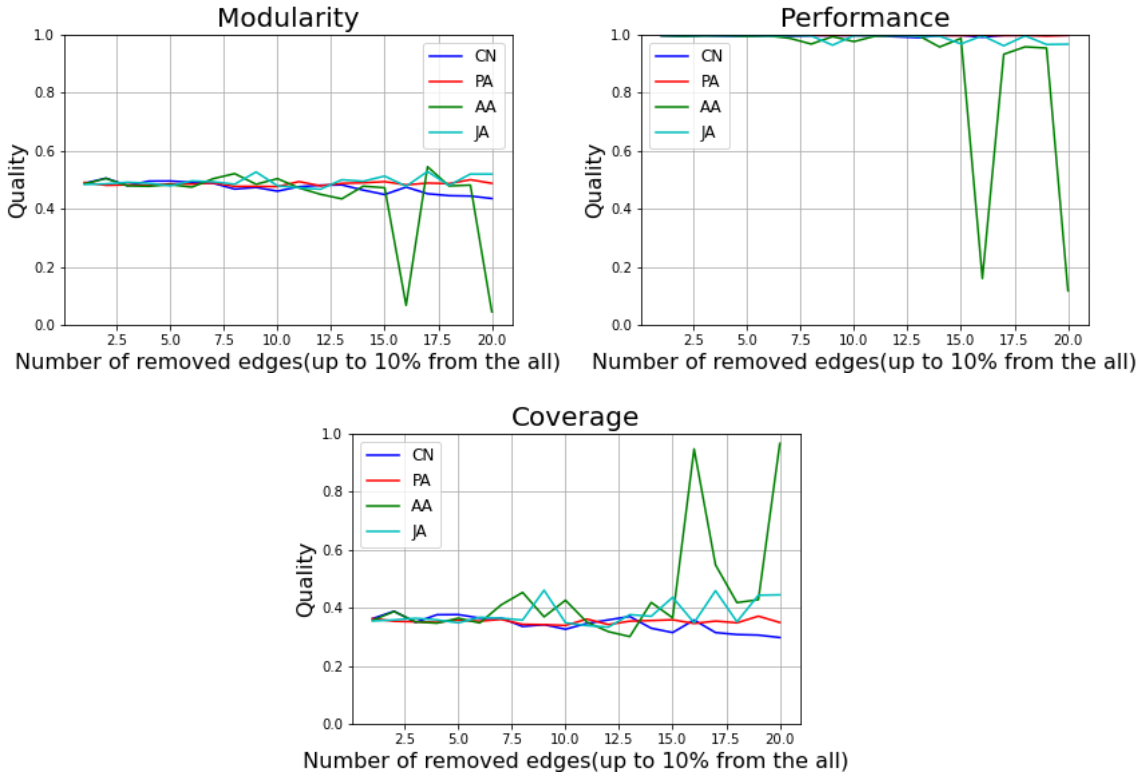


شکل ۴—۱۴ نتایج الگوریتم GMC برای مجموعه داده King James

سطح کیفیت جوامع مجموعه داده King James نسبتا پایین است. برای همین روش‌ها پیشنهادی، جوامع این مجموعه داده را خیلی خوب بهبود داده است. بر اساس معیار ماژولاریتی و پرفورمنس الگوریتم PA بهتر از همه الگوریتم‌ها جوامع را بهبود داده است.

۴-۵-۴- نتایج مجموعه داده Adolescent

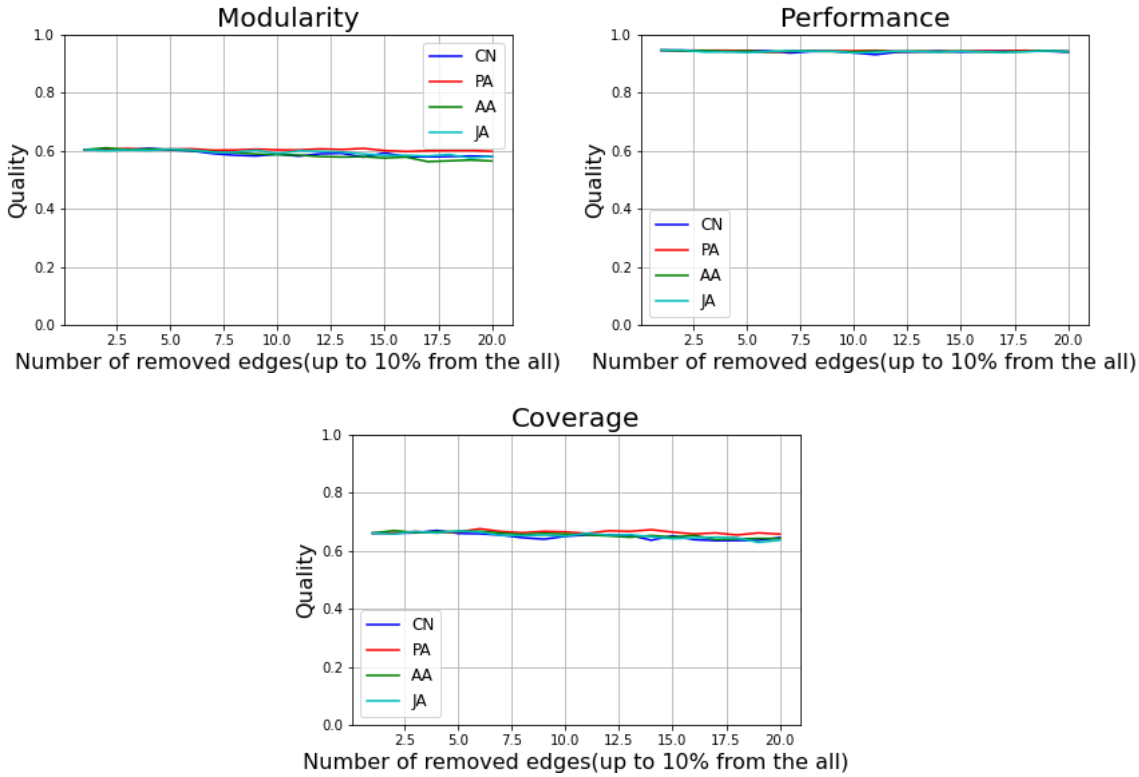
۴-۵-۴-۱- نتایج الگوریتم ALC



شکل ۴—۱۵نتایج الگوریتم ALC برای مجموعه داده Adolescent

بر اساس معیار ماژولاریتی JA بهترین عملکرد را داشته و براساس معیار پرفورمنس PA و CN بهترین عملکرد را داشته است. اما بر اساس معیار کاوریج AA، بهترین عملکرد را داشته است.

۴-۵-۴-۲- نتایج الگوریتم GMC



شکل ۴—۱۶نتایج الگوریتم GMC برای مجموعه داده Adolescent

نتیجه تشخیص جوامع الگوریتم GMC برای مجموعه داده Adolescent برای همه الگوریتم‌ها تقریبا یکسان بوده است.

۴-۶- جمع‌بندی و تفسیر روش دوم

نتایج الگوریتم ALP

- بر اساس معیار Mod و الگوریتم ALP، روش PA بهترین عملکرد را داشته است و بعد CN
- بر اساس معیار Per و الگوریتم ALP، روش PA بهترین عملکرد را داشته است و بعد CN
- بر اساس معیار Cov و الگوریتم ALP، روش PA و JC بهترین عملکرد را داشته است

نتایج الگوریتم GMC

- بر اساس معیار Mod و الگوریتم GMC، روش PA و AA بهترین عملکرد را داشته است
- بر اساس معیار Per و الگوریتم GMC، روش AA بهترین عملکرد را داشته است و بقیه نزدیک به هم بوده‌اند
- بر اساس معیار Cov و الگوریتم GMC، روش PA بهترین عملکرد را داشته است و بقیه نزدیک به هم بوده‌اند

در کل می‌توان نتیجه گرفت که الگوریتم‌های پیشگویی پیوند برای تشخیص یال‌های ناهنجار و در شبکه‌های مختلف کارایی خوبی

دارند و همچنین در بهبود جوامع و متراکم کردن آنها نیز می توان از الگوریتم‌های پیشگویی پیوند نیز استفاده کرد.

۴-۷- جمع‌بندی

در این فصل، چهار مجموعه داده برای گراف‌های بدون وزن و چهار مجموعه داده برای گراف وزن‌دار معرفی کردیم و هر دو روش پیشنهادی را روی این مجموعه داده‌ها اعمال کردیم. ما در هر دو روش پیشنهادی از چهار الگوریتم پیشگویی پیوند بدون نظارت(AA, PA, CN, JA) استفاده کرده‌ایم. در روش پیشنهادی اول نتایج هر چهار روش پیشگویی پیوند منفی(AA, PA, CN, JA) را با هشت الگوریتم دیگر مقایسه کردیم و تاثیر الگوریتم‌های پیشگویی پیوند را در تشخیص یال‌های ناهنجار در گراف‌های بدون وزن را مشاهده کردیم. در گراف‌های وزن‌درا از پیشگویی پیوند برای بهبود جوامع استفاده کرده و نتایج را ثبت و ارزیابی کردیم. روش مقایسه به این صورت بوده است: الگوریتم را روی هر مجموعه داده اجرا کرده‌ایم و با توجه به روش پیشنهادی درصدی (باتوجه با اندازه مجموعه داده) از یال‌های ناهنجار را از آن حذف کردیم و برای ارزیابی نتایج، الگوریتم‌های تشخیص جوامع را روی مجموعه داده اعمال و نتایج را ثبت کردیم. این کار را برای چهار الگورتیم پایه پیشگویی پیوند انجام دادایم. در نتیجه مشاهده کردیم که در بیشتر مواقع با حذف یال‌های ناهنجار جوامع بهتری در مجموعه داده‌ها بدست می‌آید.

فصل پنجم: جمع‌بندی و پیشنهادات

۵-۱- جمع‌بندی

در این پایان نامه روشی جدید برای تشخیص ناهنجاری یال بر اساس پیشگویی پیوند منفی در گراف‌های بدون وزن و وزن‌دار معرفی شده و از الگوریتم‌های بدون نظارت AA، PA، JC و CN برای پیشگویی یال‌های ناهنجار استفاده شده است. روشی برای تشخیص ناهنجاری یال در گراف‌های بدون وزن و روش دیگری برای تشخیص ناهنجاری یال در گراف‌های وزن‌دار را پیشنهاد و بررسی و شاهد بهبود نتایج بوده‌ایم. روش پیشنهادی اول با هشت الگوریتم مقایسه شده و در بیشتر مواقع روش پیشنهادی برتری داشته است. در روش پیشنهادی دوم، درصدی از یال‌های ناهنجار از گراف حذف شده و این باعث بهبود جوامع شده است. برای پیشگویی پیوند منفی گراف‌های وزن‌دار، و تولید دیتای نویز در گراف‌های بدون وزن روشی جدید ارائه شده است.

یکی از مهم‌ترین یافته‌های این پایان نامه حذف یال‌های ناهنجار در گراف‌های بدون وزن و وزن‌دار با استفاده از الگوریتم‌ها پیشگویی پیوند بدون ناظر بوده است. برای اثبات و ارزیابی این روش‌ها آزمایشات مختلفی انجام داده‌ایم که در فصل چهارم تشریح شده است.

۵-۲- پیشنهادات

روش‌های پیشنهادی این پایان نامه برای تشخیص ناهنجاری در انواع شبکه‌های مختلف می‌تواند موثر باشد. با این وجود فضای مناسبی برای روش‌های بهتر در این حوزه وجود دارد. در این قسمت پیشنهادهایی برای ارایه روش‌های با کارایی بهتر بیان می‌شود.

۱. همین روش را می‌توان با دیگر الگوریتم‌های پیشگویی پیوند بانظر آزمایش کرد و نتایج را بررسی کرد

۲. می‌توان از الگوریتم‌های پیشگویی پیوند منفی پیشرفته‌تر جهت پیدا کردن یال‌های ناهنجار استفاده کرد

۳. می‌توان جهت تولید یال‌های ناهنجار یا دیتای نویز برای گراف‌های وزن‌دار و بدون وزن روش‌ها موثرتری ارائه نمود

فهرست منابع

[1] Eberle, W., & Holder, L. (2007). Anomaly detection in data represented as graphs. *Intelligent Data Analysis*, *11*(6), 663-689..

[2] Hawkins, D. M. (1980). *Identification of outliers* (Vol. 11). London: Chapman and Hall.

[3] Faloutsos, C. *Large graph mining: patterns, cascades, fraud detection ,and algorithms*. in *Proceedings of the 23rd international conference on World wide web*. 2014.

[4] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, *41*(3), 1-58.

[5] Guthrie, D., Guthrie, L., Allison, B., & Wilks, Y. (2007, January). Unsupervised Anomaly Detection. In *IJCAI* (pp. 1624-1628).

[6] Noble, C. C., & Cook, D. J. (2003, August). Graph-based anomaly detection. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 631-636).

[7] Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation*, *101*(23), e215-e220.

[8] Gulbahce, N., & Lehmann, S. (2008). The art of community detection. *BioEssays*, *30*(10), 934-938.

[9] Akoglu, L., H. Tong, and D. Koutra, *Graph based anomaly detection and description: a survey.* Data Mining and Knowledge Discovery, 2015 :(۳)۲۹ .p. 626-688.

[10] Ranshous, S., et al., *Anomaly detection in dynamic networks: a survey.* Wiley Interdisciplinary Reviews: Computational Statistics, 2015. **7**(3): p. 223-247.

[11] Chandola, V., A. Banerjee, and V. Kumar, *Anomaly detection: A survey.* ACM computing surveys (CSUR), 2009. **41**(3): p. 15.

[12] Song, X., et al., *Conditional anomaly detection.* IEEE Transactions on knowledge and Data Engineering, 2007. **19**(5): p. 631-645.

[13] Fawcett, T. and F.J. Provost. *Activity Monitoring: Noticing Interesting Changes in Behavior*. in *KDD*. 1999. Citeseer.

[14] Wong, W.-K., et al. *Bayesian network anomaly pattern detection for disease outbreaks*. in *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*. 2003.

[15] Ding, Q., et al. *Intrusion as*) *anti) social communication: characterization and detection*. in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2012.

[16] Idé, T. and H. Kashima. *Eigenspace-based anomaly detection in computer systems* .in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2004.

[17] Sun, J., Xie, Y., Zhang, H., & Faloutsos, C. (2008). Less is more: Sparse graph mining with compact matrix decomposition. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *1*(1), 6-22.

[18] Bolton, R. J., & Hand, D. J. (2001). Unsupervised profiling methods for fraud detection. *Credit scoring and credit control VII*, 235-255.

[19] Phua, C., D. Alahakoon, and V.J.A.s.e.n. Lee, *Minority report in fraud detection: classification of skewed data.* 2004. **6**(1): p. 50-59.

[20] Kumar, M., R. Ghani, and Z-.S. Mei. *Data mining to predict and prevent errors in health insurance claims processing*. in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2010.

[21] McGlohon, M., et al. *Snare: a link analytic system for graph labeling and risk detection*. in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2009.

[22] Castillo, C., et al. *Know your neighbors: Web spam detection using the web topology*. in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. 2007.

[23] Ott, M., C. Cardie, and J. Hancock. *Estimating the prevalence of deception in online review communities*. in *Proceedings of the 21st international conference on World Wide Web*. 2012.

[24] Pandit, S., et al. *Netprobe: a fast and scalable system for fraud detection in online auction networks*. in *Proceedings of the 16th international conference on World Wide Web*. 2007.

[25] Abe, N., et al. *Optimizing debt collections using constrained reinforcement learning*. in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2010.

[26] Fawcett, T. and F.J. Provost. *Combining Data Mining and Machine Learning for Effective User Profiling*. in *KDD*. 1996.

[27] Akoglu, L., Tong, H., & Koutra, D. (2015). Graph based anomaly detection and description: a survey. *Data mining and knowledge discovery*, *29*(3), 626-688.

[28] Chalupsky, H. (2003, November). Unsupervised link discovery in multi-relational data via rarity analysis. In *Third IEEE International Conference on Data Mining* (pp. 171-178). IEEE.

[29] Rattigan, M.J. and D. Jensen, *The case for anomalous link discovery.* Acm Sigkdd Explorations Newsletter, 2005. **7**(2): p. 41-47.

[30] Guimerà, R. and M. Sales-Pardo, *Missing and spurious interactions and the reconstruction of complex networks.* Proceedings of the National Academy of Sciences, 2009. **106**(52): p. 22073-22078.

[31] Mitchell, C., R. Agrawal, and J. Parker. *The Effectiveness of Edge Centrality Measures for Anomaly Detection*. in *2019 IEEE International Conference on Big Data (Big Data)*. 2019. IEEE.

[32] Bhatia, S., et al. *Midas: Microcluster-Based Detector of Anomalies in Edge Streams*. in *AAAI*. 2020.

[33] Hébert-Dufresne, L., J.A. Grochow, and A. Allard, *Multi-scale structure and topological*

anomaly detection via a new network statistic: The onion decomposition. Scientific reports, 2016. **6**: p. 31708.

[34] Das, K., S. Samanta, and M. Pal, *Study on centrality measures in social networks: a survey.* Social network analysis and mining, 2018. **8**(1): p. 13.

[35] Airoldi, E.M., et al. *Mixed membership stochastic block models for relational data with application to protein-protein interactions*. in *Proceedings of the international biometrics society annual meeting*. 2006.

[36] Eronen, L. and H. Toivonen, *Biomine: predicting links between biological entities using network models of heterogeneous databases.* BMC bioinformatics, 2012. **13**(1): p. 119.

[37] Al Hasan, M., et al. *Link prediction using supervised learning*. in *SDM06: workshop on link analysis, counter-terrorism and security*. 2006.

[38] Chen, H., X. Li, and Z. Huang. *Link prediction approach to collaborative filtering*. in *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'05)*. 2005. IEEE.

[39] Clauset, A., M.E. Newman, and C. Moore, *Finding community structure in very large networks.* Physical review E, 2004. **70**(6): p. 066111.

[40] Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American society for information science and technology*, *58*(7), 1019-1031..

[41] Lü, L., & Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications*, *390*(6), 1150-1170.

[42] Pujari, M. (2015). *Link Prediction in Large-scale Complex Networks (Application to bibliographical Networks)* (Doctoral dissertation).

[43] Zhu, X.J., *Semi-supervised learning literature survey*. 2005, University of Wisconsin-Madison Department of Computer Sciences.

[44] Kashima, H., et al. *Link propagation: A fast semi-supervised learning algorithm for link prediction*. in *Proceedings of the 2009 SIAM international conference on data mining*. 2009. SIAM.

[45] Brouard, C., d'Alché-Buc, F., & Szafranski, M. (2011, June). Semi-supervised penalized output kernel regression for link prediction.in IBISC - Informatique, Biologie Intégrative et Systèmes Complexes.

[46] Sulaimany, S., Khansari, M., Zarrineh, P., Daianu, M., Jahanshad, N., Thompson, P. M., & Masoudi-Nejad, A. (2017). Predicting brain network changes in Alzheimer's disease with link prediction algorithms. *Molecular BioSystems*, *13*(4), 725-735.

[47] Harary, F., & Nash-Williams, C. S. J. (1965). On eulerian and hamiltonian graphs and line graphs. *Canadian Mathematical Bulletin*, *8*(6), 701-709.

[48] Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, *69*(2), 026113.

[49] Raghavan, U. N., Albert, R., & Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical review E*, *76*(3), 036106.

[50] Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical review E*, *70*(6), 066111.

[51] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management*, *45*(4), 427-437.

[52] Fortunato, S. (2010). Community detection in graphs. *Physics reports*, *486*(3-5), 75-174.

پیوست ۱: واژه نامه فارسی به انگلیسی

معادل فارسی	معادل لاتین
پیشگویی پیوند	Link Predecion
یال	Edge
گره	Node
ناهنجاری	Anomaly
پرت، دور افتاده	Outliers
مشاهدات ناسازگار	Discordant Observations
استثنائات	Exceptions
انحرافات	Aberrations
شگفتی‌ها	Surprises
ناهنجاری نقطه	Point Anomalies
ناهنجاری متنی	Contextual Anomalies
ناهنجاری جمعی	Collective Anomalies
ویژگی متنی	Contextual Attributes
ویژگی رفتاری	Behavioral Attributes
الکتروکاردیوگرام	Electrocardiogram
تشخیص نفوذ	Intrusion Detection
استریم	Streaming
نمیه نظارتی	Semi Supervised
بدون نظارت	Unsupervised
مبتنی بر میزان	Host Based
مبتنی بر شبکه	Network Based
تشخیص تقلب	Fraud Detection
ابزار دقیق	Instrumentation Errors

Credit Card Fraud Detection	تشخیص تقلب در کارت‌های اعتباری
Mobile Phone Fraud Detection	تشخیص تقلب در گوشی‌های موبایل
Insurance Claim Fraud Detection	تشخیص تقلب در مطالب بیمه
Medical And Public Health Anomaly Detection	تشخیص ناهنجاری پزشکی و بهداشت عمومی
Recording Errors	خطاهای ضبط
Industrial Damage Detection	تشخیص خسارت صنعتی
Motion Detection	تشخیص حرکت
Satellite Imagery	تصاویر ماهواره‌ای
Spectroscopy	طیف سنجی
Mammographic Image Analysis	تجزیه تحلیل تصاویر ماموگرافی
Video Surveillance	نظارت تصویری
Lightness	سبکی
High Dimensional	بسیار بعدی
Very Sparse	بسیار پراکننده
Fraud	تقلب
Supervised Anomaly Detection	تشخیص ناهنجاری بانظارت
Semisupervised Anomaly Detection	تشخیص ناهنجاری نمیه نظارتی
Unsupervised Anomaly Detection	تشخیص ناهنجاری بدون نظارت
Multi-Class Classification	چند کلاسه
One-Class Classification	تک کلاسه
Neural Networks	شبکه عصبی
Bayesian Networks	شبکه‌های بیزین
Support Vector Machines	ماشین بردار پشتیبان
Rule-Based	مبتنی بر قاعده
Neural Networks	شبکه‌های عصبی

Nearest Neighbor-Based	مبتنی بر نزدیک‌ترین همسایه
Using Relative Density	استفاده از چگالی نسبی
Clustering-Based	مبتنی خوشه‌بندی
Statistical Anomaly Detection Techniques	روش‌های تشخیص ناهنجاری آماری
Information Theoretic Anomaly Detection Techniques	روشهای تشخیص ناهنجاری تئوری اطلاعات
Threshold	مقدار آستانه
Anomaly Detection In Static Graphs	تشخیص ناهنجاری در گراف‌های ایستا
Static Graphs	گراف‌های ایستا
Plain (Unlabeled) Graphs	گراف‌های ساده(بدون برچسب)
Attributed (Node-/Edge-Labeled) Graphs	گراف‌های وابسته(برچسب‌دار)
Few	نادر
Different	متفاوت
Structure-Based Patterns	الگوهای مبتنی برساختار
Community-Based Patterns	الگوهای مبتنی برجامعه
Community-Based Patterns	مبتنی بر ویژگی
Proximity Based	مبتنی بر مجاورت
Graph-Centric	گراف مرکزی
Anomaly Detection In Dynamic Graphs	تشخیص ناهنجاری در گراف‌های پویا
Temporal Anomalous Pattern Detection	تشخیص الگوهای ناهنجاری زمانی
Event Detection	تشخیص رویداد
Change-Point Detection	تشخیص نقطه تغییر
Timestamps	در یک زمان
Scalability	مقیاس پذیر
Sensitivity To Structural And Contextual Changes	حساس به تغییر ساختاری و متنی یا زمینه‌ای
Mportance-Of-Change Awareness	آگاهی از تغییرات
Feature Based	مبتنی بر ویژگی

Decomposition-Based	مبتنی بر تجزیه
Community Or Clustering-Based	مبتنی بر جامعه یا خوشه‌بندی
Window-Based	مبتنی بر پنجره
Degree Distribution	توزیع درجه
Diameter	قطر
Eigenvalues	مقادیر ویژه
Feature Extraction	استخراج ویژگی
Classification	خوشه‌بندی
Katz	کاتز
Random Walk	قدم زنی تصادفی
Random Walk	کوتاه‌ترین مسیر
Common Neighbors	همسایه مشترک
Preferential Attachment	پیوست امتیازی
Adamic-Adar	آدامیک-آدر
Jaccard	جاکارد
Resource Allocation	اندیس تخصیص منبع
Positive Link Prediction	پیشگویی پیوند مثبت
Negative Link Prediction	پیشگویی پیوند منفی
Minimum Description Length Principle	حداقل طول توصیف
Link Mining	لینک کاوی
Structural Based	مبتنی بر ساختار
Behavioral Based	مبتنی بر رفتار
Precision	دقت
Recall	بازیابی
F-Measures	معیار اف
Asynchronous Label Propagation Algorithm	الگوریتم انتشار برچسب نامتقارن

Label Propagation Algorithm	الگوریتم انتشار برچسب
Graph Line	لاین گراف

Abstract

Anomaly detection is a very important and vital task, and it has many applications in various fields such as security, health, finance, health care and law enforcement. In recent years, many techniques for diagnosing abnormalities or discarded data have been proposed in non-structured sets of multidimensional data, some of which have focused on graph structure. This thesis is based on the detection of edge anomaly in the graph. Two methods have been proposed based on negative link prediction for the detection of edge anomalies. The first methods are for simple graphs and the second one is for weighted graphs. In both methods anomalous edges are removed from the graph based on the negative link prediction algorithm. For both proposed methods, four unsupervised basic link prediction algorithms, the Jaccard's Coefficient, Preferential Attachment, Common Neighbors, Adamic/Ader were used. We also used four standard dataset Dolphins, Trinity100, Netscience and Email networks for simple graphs, and four Lesmis, King James, Netscience and Adolescent networks for weighted graphs. In order to evaluate the first proposed method, we added some anomaly nodes to the graph and employed eight extra competitor meaures in order to find them. The evaluation metric was precision, recall and f-measure. For the second method, we got help from community detection improvement for evaluation. Two community detection algorithms, Asynchronous Label Propagation and optimized Girvan-Newman were used. Evalution metric for quality of the community detection methods was modularity, performance and coverage. We also proposed a new approach for weighted negative link predictionand a new method for adding the anomolus links into weighted graphs. Results show the superiority of the proposed methods. Future works can deploy other link prediction algorithms and other graph types.

Keyword: Anomaly Detection, Anomaly Detection edge, Anomlay Detection Graph, Link prediction, Negative link prediction

University of Kurdistan

Faculty of Engineering

Department of Computer Software Engineering

Title:

A New Edge Anomaly Detection Based On Negative Link Prediction

By:

Arman Hossiny

Supervisor:

Dr. Fardin Akhlaghian

Dr. Sadegh Sulaimany

June, 2020