
بسم الله الرحمن الرحیم

دانشکده مهندسی کامپیوتر و فناوری اطلاعات

پایان‌نامه کارشناسی ارشد هوش مصنوعی و رباتیکز

تشخیص تقلب در تراکنش‌های مالی با استفاده از رویکردهای یادگیری عمیق

نگارنده: سید علی اکبر حسینی

استاد راهنما:

دکتر هدی مشایخی

اسفند ماه ۱۴۰۰

در این صفحه صورت جلسه دفاع را قرار دهید. لازم است پس از صحافی این صفحه مجدداً توسط دانشکده مهر گردد و استاد راهنما با امضای خود اصلاحات پایان‌نامه را تایید کند.

تقدیم اثر

تقدیم به خانواده عزیزم

بخاطر عشق بی‌پایان و حمایت‌های‌شان

تشکر و قدردانی

با تشکر از استاد راهنمای گرانقدر سرکار خانم دکتر هدی مشایخی که در تمام طول مسیر این پژوهش کمک‌های بسیاری نموده‌اند.

تعهدنامه

اینجانب **سید علی اکبر حسینی** دانشجوی دوره کارشناسی ارشد (دکتری) رشته **مهندسی کامپیوتر گرایش هوش مصنوعی و رباتکیز** دانشکده مهندسی کامپیوتر و فناوری اطلاعات دانشگاه صنعتی شاهرود نویسنده پایان‌نامه **تشخیص تقلب در تراکنش‌های مالی با استفاده از رویکردهای یادگیری عمیق** تحت راهنمایی **دکتر هدی مشایخی** متعهد می‌شوم.

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است .
- در استفاده از نتایج پژوهش‌های محققان دیگر به مرجع مورد استفاده استناد شده است .
- مطالب مندرج در پایان‌نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است .
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود است و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « Shahrood University of Technology » به چاپ خواهد رسید .
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان‌نامه تأثیرگذار بوده‌اند در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- در کلیه مراحل انجام این پایان‌نامه ، در مواردی که از موجود زنده (یا بافت‌های آن‌ها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است .
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است .

تاریخ: اسفند ماه ۱۴۰۰

امضای دانشجو: سید علی اکبر حسینی

چکیده

شناسایی رفتار و الگوی متفاوت داده‌ها در میان حجم زیاد اطلاعات موجود، از اهمیت زیادی برخوردار است. یکی از مهم‌ترین حوزه‌های حال حاضر درمورد تشخیص الگوی متفاوت و ناهنجاری در داده‌ها، حوزه‌های مالی و بخش معاملات است. به دلیل حجم بالای تراکنش‌ها، تعداد زیاد سرمایه‌گذاران و پیچیدگی فرآیند انجام معاملات، تقلب و تبانی اقتصادی از جمله معضلاتی است که در این بخش وجود دارد. با استقرار و رشد دولت الکترونیک در بسیاری از کشورها، بسیاری از مناقصات و مزایده‌های انجام شده در بخش دولتی از روش‌های سنتی خارج شده و به صورتی الکترونیکی انجام می‌شود. پژوهش حاضر با هدف شناسایی افراد مشکوک در مناقصات بخش دولتی انجام شد و به بررسی تاثیر استفاده از روش‌های یادگیری عمیق بازنمایی گراف در شناسایی رفتار مشکوک داده‌ها پرداخته است. در این پژوهش با کمک روش خوشه‌بندی به تحلیل رفتار مشتریان در معاملات پرداخته شد و الگوی داده‌ها در خوشه‌های مختلف تحلیل شد.

مجموعه دادگان استفاده شده در پژوهش حاضر، شامل مناقصات انجام شده توسط دستگاه‌های دولتی در کشور ایران است که در بازه‌ی زمانی ۱۳۸۹ تا ۱۳۹۸ در سامانه تدارک دولتی (ستاد) در دسترس عموم قرار دارد. بعد از مدل‌سازی و تبدیل داده‌ها در قالب گراف معاملات، با روش‌های تحلیل شبکه‌های گرافی شاخص‌های مهم گراف استخراج شد و به استخراج انجمن‌های پرارزش گراف معاملاتی پرداخته شد و تامین‌کنندگان و دستگاه‌های اجرایی تاثیرگذار در معاملات بزرگ شناسایی شد. برای اجرای الگوریتم خوشه‌بندی و تحلیل خوشه‌ها بر روی داده‌های انجمن‌ها، با استفاده از رویکردهای یادگیری عمیق بازنمایی گراف، تعبیه گره‌های گراف صورت گرفت. ویژگی‌های هر گره معاملاتی (ارزش ریالی معاملات و تعداد معاملات انجام شده) به همراه ویژگی‌های بدست آمده از طریق تحلیل گراف و بردار ویژگی تعبیه گراف، به عنوان ورودی داده‌های الگوریتم خوشه‌بندی در نظر گرفته شد. و در نهایت با تحلیل خوشه‌ها و داده‌های درون هریک از آن خوشه‌ها داده‌های مشکوک و با رفتار متفاوت شناسایی شد.

یافته‌های تحقیق نشان می‌دهند که استفاده از ویژگی‌های گراف به همراه بردار ویژگی تعبیه شده گره‌های گراف که توسط روش‌های یادگیری عمیق بازنمایی گراف به دست آمده، در خوشه‌بندی بهتر و با کیفیت‌تر داده‌ها کمک قابل توجهی داشتند. بعلاوه، این پژوهش نشان داد که هرچه طول بردار ویژگی در تعبیه گراف بیشتر باشد، به دلیل در دسترس بودن اطلاعات بیشتر گره‌های همسایه در گراف، باعث عملکرد بهتر در خوشه‌بندی داده‌ها می‌شود و شناسایی داده با رفتار مشکوک در بین داده‌ها با دقت بالاتری صورت می‌گیرد.

کلمات کلیدی: تشخیص ناهنجاری، دولت الکترونیک، گراف‌کاوی، تعبیه گراف، یادگیری بازنمایی گراف، خوشه‌بندی، الگوریتم **K-means**، یادگیری عمیق

فهرست مطالب

فهرست جداول

فهرست اشکال

فصل۱ مقدمه

۱-۱ تشخیص ناهنجاری

یکی از مهم‌ترین مسائل در دنیای امروز که رشد روزافزون تعداد داده‌ها را در آن شاهد هستیم، تشخیص ناهنجاری در طیف وسیع داده‌ها در حوزه‌های مختلف است. مفهوم داده‌های ناهنجار اولین بار در سال ۱۹۶۹ توسط Grubb به وجود آمد. تعریف دقیقی از تشخیص ناهنجاری وجود ندارد، اما به طور کل تشخیص ناهنجاری مسئله‌ی یافتن داده‌هایی است که با رفتار معمولی دیگر داده‌ها مطابقت ندارند. در حوزه‌های مختلف دلایل مختلفی برای وجود چنین داده‌های ناهنجاری وجود دارد. خطاهای انسانی در جمع‌آوری و وارد کردن داده‌ها، تغییر غیرمنتظره در رفتار و عملکرد سامانه، خطاهای مربوط به ابزارهای مکانیکی، رفتار متقلبانه و مشکوک افراد در سامانه، فعالیت‌های مخربانه در شبکه‌های کامپیوتری، خطاهای مربوط به ابزارهای مکانیکی در محیط‌های صنعتی، خطاهای نمونه‌برداری و تغییرات محیطی (مانند دما) از دلایلی هستند که باعث به وجود آمدن داده‌های ناهنجار می‌شوند. داده پرت[۱]، شناسایی نو و جدید بودن داده[۲]، اختلالات داده‌ها[۳]، داده‌های عجیب و نامأنوس[۴] از دیگر نام‌هایی است که متناسب با حوزه‌های کاربردی مختلف به داده‌های ناهنجار اطلاق می‌شود[۱].

مسئله تشخیص ناهنجاری را می‌توان از جنبه‌های مختلف مورد بحث و بررسی قرار داد. یکی از مهم‌ترین ابعاد هر روش تشخیص ناهنجاری، دانستن ماهیت داده‌های ورودی آن است. هر نمونه داده با مجموعه‌ای از ویژگی‌ها توصیف می‌شود. داده‌های ترتیبی[۵] (صوت، متن، سری‌های زمانی) و غیر ترتیبی (تصویر، رکوردهای عددی و دیگر داده‌ها)، دادهای دودویی، پیوسته، طبقه‌ای، تک متغیره و چند متغیره از انواع داده‌های ورودی هستند که مشخص‌کننده کاربرد روش‌های متفاوت تشخیص ناهنجاری‌اند.

[1] Outlier Data
[2] Novelty Detection
[3] Abnormality in Data
[4] Unexpected Data
[5] Sequential Data

جنبه مهم دیگر، نوع بی‌نظمی موجود در داده‌هاست. بی‌نظمی نقطه‌ای[1]، مفهومی[2] و تجمعی[3] از انواع بی‌نظمی‌هایی هستند که در داده‌های حوزه‌های مختلف شناسایی می‌شوند. مسئله‌ی مهم دیگر در تشخیص ناهنجاری، وجود یا عدم وجود داده‌های برچسب‌دار است. روش‌های متفاوتی در تشخیص ناهنجاری وجود دارد؛ الگوریتم‌های یادگیری بدون نظارت[4] (داده‌هایی که از وضعیت ناهنجاری آن مطلع نیستیم)، الگوریتم‌های یادگیری با نظارت[5] (داده‌هایی که برچسب ناهنجاری و یا داده معمولی دارند) و روش نیمه نظارتی[6] که از هرکدام از این روش‌ها متناسب با نوع داده موجود و کاربرد آن استفاده می‌شود. و درنهایت، خروجی یک الگوریتم و روش تشخیص ناهنجاری می‌تواند به صورت امتیاز و یا به صورت برچسب (داده هنجار و داده ناهنجار) باشد.

با توجه به مفهوم داده‌های ناهنجار و جنس داده‌های ورودی، ناهنجاری با بسیاری از برنامه‌های کاربردی دنیای واقعی مرتبط هستند. برنامه‌هایی مثل نفوذ در شبکه[7]، کلاهبرداری و تقلب در تراکنش‌های مالی و بانکی[8]، تقلب در مسائل مربوط به بیمه، ناهنجاری در داده‌های حوزه سلامت و پزشکی، داده‌های پرت با توجه عیوب در صنعت و ماشین‌آلات صنعتی، پردازش تصویر و ویدئو و اینترنت اشیاء از مواردی در دنیای واقعی هستند که امکان وجود داده ناهنجار در آن‌ها زیاد است. به عنوان مثال در زمینه پزشکی وجود ناهنجاری در تصاویر اسکن MRI مغزی نشان‌دهنده تومور بدخیم می‌تواند باشد[۲]. به طور مشابه، در حوزه‌های مالی نیز رفتار مشکوک در تراکنش‌های کارت اعتباری منجر به اعمال مخربانه می‌شود[۳]. و در شبکه‌های کامپیوتری نیز الگوی ترافیکی غیرمعمول یک شبکه نشان‌دهنده هک شدن یا مورد حمله قرار گرفتن شبکه نیز است[۴].

[1] Point Anomaly
[2] Contextual Anomaly
[3] Collective Anomaly
[4] Unsupervised Learning
[5] Supervised Learning
[6] Semi-Supervised Learnin
[7] Network Intrusion
[8] Financial Fraud

۱-۲ چالش‌های تشخیص ناهنجاری

در راستای تشخیص ناهنجاری در داده‌ها، چالش‌هایی وجود دارد که این مسئله را دشوار می‌کند. یکی از چالش‌ها شناسایی و تعیین حد و مرز بین داده‌های پَرت و داده‌های نویز است. داده پرت داده‌ای است که خارج از محدوده‌ی مورد انتظار ما در مجموعه داده‌ها است. ولی داده‌ی نویز رفتار اشتباه و غلطی دارد. تعداد کم داده‌های ناهنجار و عدم تعادل از نظر تعداد بین داده‌های معمولی و داده ناهنجار در یک مجموعه داده از دیگر مسائلی است که باعث می‌شود دقت رویکردهای تشخیص داده ناهنجار کاهش یابد. از دیگر چالش‌هایی که محققان بسیاری را به تحقیق بیشتر و ارائه رویکردهای متفاوت در این حوزه علاقه‌مند کرده، گستردگی در حوزه‌های مختلف و اهمیت شناسایی داده‌ی ناهنجار است که باعث می‌شود تکنیک‌های و رویکردهای شناسایی مختص همان حوزه باشد و برخی از الگوریتم‌های و تکنیک‌ها را نتوان در حوزه‌های دیگر استفاده کرد. و درنهایت در دسترس نبودن داده برچسب‌گذاری شده نیز از چالش‌ها بحث‌برانگیز در حوزه است.

کسب اطلاعات مفید و تصمیم‌گیری بهتر در مورد داده‌ها از تاثیرات مثبت شناسایی داده‌ها ناهنجار است. بدین معنی که با فهمیدن و درک در مورد روند و رفتار داده ناهنجار و تمیز دادن این نوع داده از داده‌های با رفتار معمولی می‌توان در حوزه‌های مختلف باعث جلوگیری از افشای اطلاعات حساس، جلوگیری از دسترسی‌های غیرمجاز و پیش‌گیری از تصمیم‌گیری‌های خطا شد.

۱-۳ بیان مسئله و ضرورت انجام تحقیق

همان‌طور که گفته شد در بسیاری از موارد، ناهنجاری می‌تواند باعث تأثیرات جدی‌ای در حوزه‌های مختلف بشود. نکته‌ی حائی اهمیت این است که تشخیص روند داده‌های ناهنجار در داده‌ها با توجه به نظر و تصمیم تحلیلگران داده و متخصصان در آن حوزه ایجاد می‌شود. تشخیص تقلب درحوزه‌های مالی یکی از کاربردی‌ترین و مهم‌ترین حوزه‌های روش‌های تشخیص ناهنجاری است و در این حوزه کارهای زیادی

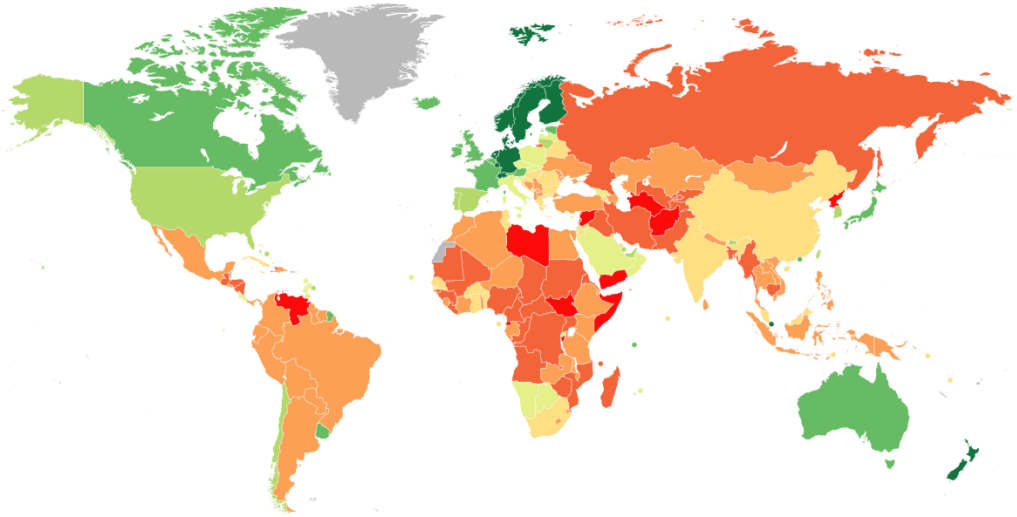
در بخش تحقیقات دانشگاهی و در دنیای واقعی صورت گرفته است. شناسایی و کشف تقلب‌های مالی مانند تقلب‌های بانکی (پول‌شویی[1]، تقلب در کارت‌های اعتباری[2] و تقلب در مسائل مربوط به وام)، تقلب‌های مربوط به مسائل بیمه (تقلب در مسائل بیمه حوزه سلامت و بیمه اتومبیل)، تقلب‌های شرکتی یا گروهی (تقلب در صورت‌حساب‌های مالی[3] و تقلب در مسائل مربوط به مناقصه و مزایده‌ها[4]) و اخیراً شناسایی رفتار مشکوک افراد در حوزه رمز ارزها، از مصادیق تقلب‌های مالی و رفتار مشکوک هست که از اهمیت زیادی برخوردارند و ضررهای که در این بخش متوجه بانک‌ها، دولت‌ها و شرکت‌ها می‌شود نگرانی‌های بسیاری را به وجود آورده است.

در طی دو دهه‌ی اخیر شاهد پیشرفت و رونق بسیار زیادی در زمینه دولت الکترونیک بوده‌ایم. پیشرفت سریع کشورهای توسعه‌یافته و حرکت رو به رشد کشورهای در حال پیشرفت از روش‌های سنتی بانکی و خریدوفروش به سوی دولت الکترونیک[5] بسیار چشمگیر بوده است و همچنان در حال افزایش است. این گذر از عصر سنتی به سوی استفاده از فناوری‌های اطلاعات و ارتباطاتی دلایلی همچون افزایش تقاضای شهروندان به منظور ارائه خدمات سریع‌تر و کاراتر، افزایش فشارهای سیاسی و ویژگی‌های بسیار خوب فناوری‌های اطلاعاتی را دارا است که دولت‌ها را به این سمت کشانده است [۵].

ازجمله معضلات اقتصادی که کشورها (بخصوص کشورهای در حال توسعه) از قدیم تا به حال با آن درگیر بوده‌اند مسئله فساد اقتصادی در بخش معاملاتی است که بخش دولتی متولی انجام آن است. دولت‌ها در هر کشور مبالغ بسیار زیادی را برای خرید کالا، خدمات و توسعه زیرساخت در کشور هزینه می‌کند که این امر از بستر دولت الکترونیک و بیرون‌سپاری به شرکت‌ها انجام می‌شود. فساد اقتصادی به معنی نقض قوانین موجود و استفاده از اموال دولتی برای تأمین منافع و سود شخصی خود است. با توسعه بیش‌ازپیش دولت الکترونیک که در آن معاملات در بستر فناوری اطلاعات و در سامانه‌ای الکترونیکی

[1] Money Laundering
[2] Credit Card Fraud
[3] Financial Statement Fraud
[4] Procurement Fraud
[5] E-Government

انجام می‌شود، نیاز دولت‌ها به شناسایی متخلفان و متقلبان در راستای انجام الکترونیکی تامین کالا و مناقصات و مزایده‌ها به امری بسیار مهمی تبدیل گردیده است. در شکل ۱-۳ شاخص ادراک فساد[1] را مشاهده می‌کنیم که نشان‌دهنده‌ی میزان فساد و تبانی در بخش دولتی خریدوفروش شرکت‌ها در کشورها را نشان می‌دهد. این داده‌ها که از طریق یک سازمان مستقل و غیردولتی به نام Transparency International در دسترس است عددی بین صفر (بیشترین میزان فساد) تا ۱۰۰(خالی از فساد) است. رنگ سبز نشان‌دهنده کمتر بودن فساد در آن کشورها، و رنگ قرمز نشان‌دهنده‌ی فساد بیش‌ازحد در آن کشور است.



شکل ۱-۱: شاخص CPI برای شناسایی تبانی در بخش عمومی و دولتی در کشورها[2] (بر اساس داده‌های ویکی‌پدیا در سال ۲۰۲۱)

دولت الکترونیک تغییرات اساسی‌ای را در الگوی سنتی و پیشین دولت‌ها ایجاد کرده و باعث از بین رفتن سلسله‌مراتب طولانی، کاهش هزینه‌های دولتی، کاهش اتلاف وقت به دلیل پراکندگی سازمان‌های دولتی، شفافیت اطلاعات مالی و کاهش تخلفات و فساد اداری شده است. یکی از ویژگی‌های حیاتی سامانه‌های دولت الکترونیک (سامانه‌های یکپارچه معاملات دولتی) نظارت کامل و دقیق بر روند اجرای معاملات از ابتدا تا انتها است و کنترل‌هایی هم از بُعد مقدار و از هم بُعد مبلغ و هزینه‌های انجام شده، بر

[1] Corruption Perceptions Index (CPI)
[2] https://en.wikipedia.org/wiki/Corruption_Perceptions_Index

روی داده‌ها اعمال می‌شود و روند اطلاعات مورد بررسی قرار می‌گیرد[۶].

دولت الکترونیک، استفاده دولت و سایر سازمان‌های دولتی از فناوری اطلاعات، به منظور ایجاد تحول در رابطه با شهروندان، مراکز تجاری و سایر مواردی است که با دولت در حال تعامل هستند و به افراد، تسهیلات لازم جهت دسترسی مناسب به اطلاعات و خدمات دولتی، اصلاح کیفیت آن‌ها و ارائه فرصت‌های گسترده برای مشارکت در فرآیندها و نهادهای مردم‌سالار را می‌دهد[۷].

امروزه در عصر فناوری اطلاعات دولت‌ها دریافته‌اند که کاربردهای استفاده از فناوری‌های اطلاعاتی در بخش دولتی و پیاده‌سازی دولت الکترونیک امری اجتناب‌ناپذیر است و جهت همگام شدن با تغییرات در دنیا و کشورها و آسان‌سازی خدمات‌رسانی به شهروندان باید آن را به کار گرفت. حجم معاملات و خریدوفروش دولت در کشورهای توسعه‌نیافته که زیرساخت‌های اقتصادی آن مؤلفه‌های رشد و استحکام لازم را ندارند تا بخش خصوصی قدرتمند تشکیل گردد، در مقایسه با معاملات بخش غیردولتی بسیار زیاد است. همین ویژگی میزان تبانی در معاملات را افزایش می‌دهد.

در حوزه معاملات دولتی و دولت الکترونیک، فساد، مصادیق مختلفی می‌تواند داشته باشد؛ تبانی[1]، تقلب، پول‌شویی، کسب نامشروع، اعمال‌نفوذ و تجارت مبتنی بر پول. خرید کالا و خدمات به وسیله‌ی دولت‌ها و نهادهای دولتی بخش بزرگی از اقتصاد ملی کشورها را تشکیل می‌دهد. خریدوفروش و مناقصات بخش دولتی به دلیل حجم و ارزش بالای آن یکی از بخش‌های مهم یک اقتصاد به شمار می‌آیند و به همین دلیل فساد اقتصادی و تقلب در این بخش می‌تواند باعث آسیب اقتصادی بزرگی بشود[۸].

فرآیند معاملات دولتی به دلیل سطح بالای هزینه‌های انجام شده در معاملات، حجم تراکنش‌ها، پیچیدگی فرآیند انجام معاملات و تعداد بالای سرمایه‌گذاران در آن، همیشه با احتمال ارتکاب اعمال غیرقانونی همراه است. فساد مالی مسئله‌ای است فراگیر و جهانی که امروزه بسیاری از کشورها و به‌خصوص کشورهای در حال توسعه با آن دست به گریبان‌اند. بخش دولتی به لحاظ انجام امور و انعقاد

[1] Collusion

قراردادهای دولتی در راستای تأمین کالاها و خدمات عمومی، بیشتر در معرض فسادهای مالی در حوزه معاملات دولتی و قرارداد قرار دارد. فساد مالی مفهومی کلی بوده و دارای مصادیق متعدد است[۹].

شکل ۱-۲: دلایل اهمیت شناسایی تقلب در معاملات دولتی

امروزه ماندگاری دولت الکترونیک، بدون بهره‌مندی از کاربردهای مفید فناوری اطلاعات قابل‌تصور نیست، اما از طرفی به‌کارگیری فناوری اطلاعات می‌تواند چالش‌های دیگری را برای استفاده‌کننده‌گان فراهم آورد و موجبات آسیب‌پذیری هرچه بیشتر را آن‌ها را فراهم کند.

در این پژوهش تلاش می‌شود تا با استفاده از مدل‌سازی و تحلیل گراف، رویکردهای یادگیری عمیق و تکنیک‌های یادگیری ماشین به شناسایی افراد مشکوک و تبانی‌ها در داده‌های حوزه معاملات دولت (که به صورت الکترونیک و در بستر ایجاد شده توسط دولت انجام می‌شود) بپردازیم. در این تحقیق که از داده‌های سامانه تدارکات دولت[۱] در کشور ایران استفاده‌شده است، با تعیین الگو و استخراج ویژگی‌های افراد حاضر در معاملات به تعیین افراد مشکوک در گراف معاملات مناقصات می‌پردازیم.

۱-۴ چالش‌های مسئله

تقلب و تبانی به‌عنوان پدیده‌ای نامطلوب در صنعت دولت الکترونیک شناخته می‌شود. علیرغم تلاش‌های فراوانی که برای مبارزه و جلوگیری از پدیده تقلب و تبانی صورت گرفته، به دلیل تغییر رفتار معامله‌گران در سامانه‌های خریدوفروش و مناقصات، مسئله کشف و پیشگیری کامل از تقلب و تبانی امکان‌پذیر نیست[۱۰]. از دیگر چالش‌های موجود، عدم پویایی رویکردهای موجود است که موجب

[۱] سامانه تدارکات دولت (ستاد) به آدرس اینترنتی: https://setadiran.ir/

می‌شود این روش‌ها برای شناسایی رفتار مشکوک مناسب نباشد و در مقابل روش‌های جدید تقلب کارساز نباشند. بدین روی همچنان باید در حال تغییر و بهبود روش‌های یادگیری ماشین و داده‌کاوی برای رسیدن به الگوی تغییر یافته داده‌ها بود.

از دیگر چالش‌های این مسئله عبارت‌اند از[۱۱]:

- ✓ کاهش دقت شناسایی رفتار متقلبانه به دلیل عدم تعادل بین توزیع داده‌ها (داده‌های متقلبانه و غیر متقلبانه)
- ✓ ابعاد بالای فضای داده‌های ورودی و ویژگی‌های آن‌ها
- ✓ تغییر مداوم رفتارها و الگوهای متقلبانه
- ✓ عدم وجود داده برچسب‌دار
- ✓ هزینه به‌کارگیری روش‌های مختلف
- ✓ پاسخگویی سریع و برخط و شناسایی رفتار متقلب

باوجوداینکه محققین در این زمینه تلاش‌های بسیار را انجام داده‌اند، اما از موانع اساسی برای گسترش روش‌های شناسایی تقلب و فساد در حوزه‌های مالی و سامانه‌های خریدوفروش بخش دولتی و مناقصات و مزایده‌ها، عدم تبادل دانش در این زمینه بین کشورها و شرکت‌ها است. در این حوزه به دلیل مسائلی همچون محرمانگی و امنیت داده و مسائل رقابتی موجود، تعداد کارهایی که قابلیت عملیاتی شدن را داشته باشند بسیار محدود و اندک هستند و تبادل اطلاعات در بین روش‌ها منحصرا در حوزه تحقیقات تئوری می‌باشد. در حوزه‌های مالی به دلیل نبودن داده برچسب‌دار، توجه ویژه‌ای به روش‌های تشخیص بدون نظارت (خوشه‌بندی) شده است[۱۲].همچنین امروزه به دلیل حجم زیاد داده‌ها و ماهیت رابطه‌ای بودن داده‌ها، تحلیل داده‌ها و مدل‌سازی گرافی داده‌ها از چالش‌های مهمی است که تحقیقات روزافزونی در این مورد در حال انجام است.

۱-۵ کاربرد انجام تحقیق

استفاده از فناوری اطلاعات و ارتباطات، از پارامترهای مؤثر در شناسایی و جلوگیری از فساد اقتصادی است که در دهه‌ی اخیر مورد توجه محققان و شرکت‌ها و دولت‌ها قرار گرفته است. در گذشته که معاملات و خریدوفروش‌های در دولت الکترونیک به صورت دستی و به روش سنتی انجام می‌شده، تشخیص و شناسایی افرادی که فعالیت مخربانه در فرآیند انجام معاملات داشتند آسان نبوده، ولی امروزه با توجه به گسترش بیش‌ازپیش استفاده از اینترنت و فناوری اطلاعات و ارتباطات، شناسایی این‌گونه افراد در سیستم یکپارچه معاملات دولتی که در لوای دولت الکترونیک است تسهیل گردیده است.

تشخیص و حذف افراد مشکوک از معاملات دولتی باعث حفظ و ارتقای سلامت سامانه معاملات دولتی می‌شود. از دیگر کاربردهای انجام چنین پژوهش عبارت‌اند از:

- کسب شفافیت در رویه‌های انجام معاملات؛
- افزایش انصاف و اعتماد در فرآیند معاملات؛
- افزایش رقابت میان تأمین‌کنندگان کالا و خدمات؛
- افزایش کارایی و صرفه اقتصادی در معاملات دولتی؛
- فراهم کردن رفتار برابر و منصفانه با تأمین‌کنندگان کالا و خدمات؛
- ترغیب و مشارکت بخش خصوصی در فرآیند معاملات[۱۳]

۱-۶ نوآوری روش ارائه‌شده

در این پژوهش با استفاده از رویکردهای مختلف و ترکیب آن‌ها به شناسایی و کشف رفتار مشکوک داده‌ها پرداختیم. به دلیل ماهیت رابطه‌ای بودن داده‌ها و شکل گرافی داده‌ها در وهله‌ی اول به مدل‌سازی گرافی داده‌ها پرداختیم. سپس با استفاده از تکنیک‌های گراف‌کاوی اطلاعات مفیدی را از مدل گرافی معاملات به دست آوردیم. در مرحله‌ی بعد با استفاده از روش‌های یادگیری بازنمایی در گراف، به تعبیه نودهای گراف پرداختیم و برداری از ویژگی‌های هر گره را به دست آوردیم که اطلاعات دیگر گره‌ها و

همسایگان در گراف را داراست و سپس الگوریتم داده‌کاوی‌ای را بر روی ویژگی‌های داده‌های گراف و ویژگی‌های به دست آمده از تکنیک‌های گراف‌کاوی و بردار ویژگی به دست آمده از روش‌های یادگیری عمیق اعمال کردیم. در آخر با اجرای مراحل خوشه‌بندی و تحلیل خوشه‌ها و کیفیت آن‌ها به شناسایی رفتار مشکوک و الگوی نامتعارف داده‌ها پرداختیم. نوآوری اصلی پژوهش حال حاضر مدل‌سازی داده‌ها در قالب گرافی و استفاده از روش‌های یادگیری بازنمایی گراف در شناسایی ناهنجاری است.

۱-۷ سؤالات و فرضیه‌ها مسئله

هر تحقیق در پاسخ به مشکل یا مسئله‌ای مطرح می‌شود، این امر سؤالات زیادی را در ذهن پژوهشگر ایجاد می‌کند که همین امر موجب پیدایش فرضیاتی می‌شود. پژوهشگر با جمع‌آوری اطلاعات و آمار مورد نیاز به تجزیه و تحلیل آن‌ها می پردازد و از این طریق به سؤالات پژوهشی پاسخ گفته و به تأیید یا رد فرضیات می پردازد. سوالاتی که در این تحقیق برای پژوهشگر و خوانندگان مطرح می‌شود به شرح زیر است:

الگوی رفتار مخربانه و مشکوک در تراکنش‌های مالی و خریدوفروش و معاملات و مناقصه‌ها چطور تعریف می‌شود؟ چه حد آستانه‌ای باید برای مخرب بودن و مشکوک بودن رفتار فرد مخرب تصور کرد؟

معیارها و کارایی یک سیستم تشخیص تقلب به چه صورت باید باشد تا با دقت بالا به شناسایی رفتار مخربانه در تراکنش‌ها منجر شود؟

آیا حجم و تعداد تراکنش‌ها و معاملات در شناسایی دقیق‌تر الگو و رفتار مخربانه مؤثر است؟ بدین معنی که با تعداد داده بیشتر به نتیجه بهتر می‌توان رسید یا خیر؟

چه ویژگی‌هایی از تراکنش‌های مالی و خرید فروش هستند که از اهمیت بیشتری برخوردارند و می‌توان بهتر الگوهای متقلبانه افراد را با استفاده شناسایی کرد؟

فرضیات موجود در مورد داده‌های ناهنجار:

- داده‌های ناهنجار در گروه‌هایی دورتر از دیگر گروه‌ها هستند.

- در یک گروه از داده‌ها، داده‌های ناهنجار از مرکز گروه دورتر هستند.
- هر چه امتیاز پرت بودن یک داده بیشتر باشد، آن داده به عنوان داده ناهنجار در نظر گرفته می‌شود.
- ایجاد یک حد آستانه برای امتیاز دادن به پرت بودن داده با توجه به حوزه کاربردی و یک فرد تحلیلگر داده انجام می‌شود.

۱-۸ جمع‌بندی

در این فصل ابتدا به تعریف مختصری از تشخیص ناهنجاری پرداختیم و داده ناهنجار را تعریف کردیم. سپس مسئله تشخیص تقلب در تراکنش‌های مالی را از جنبه‌های مختلف بررسی کردیم. چالش‌های موجود را برشمردیم و از اهمیت و کاربردهای تحقیق صحبت نمودیم. سؤالات و فرضیه‌هایی را مطرح کردیم که درصدد پاسخگویی به آن‌ها در طول این پژوهش نیز هستیم. در ادامه به مرور ادبیات پیشین و کارهای انجام شده در این حوزه می‌پردازیم و سپس به توضیح راهکار پیشنهادی و نتایج و ارزیابی می‌رسیم.

فصل ۲ مروری بر پژوهش‌های پیشین

۲-۱ مقدمه

همان‌طور که در فصل پیش نیز به آن اشاره کردیم، تشخیص تقلب در معاملات بخش دولتی و مناقصات انجام شده در این بخش، از موضوعات حائز اهمیت هم در بین تحقیقات دانشگاهی و محققین بوده و هم در دنیای واقعی بسیاری از شرکت‌ها و دولت‌ها به دنبال به ایجاد یک سامانه دقیق و یکپارچه جهت کنترل و مدیریت این امر مهم هستند. در دو دهه‌ی گذشته تحقیقات بسیار زیادی در این حوزه صورت گرفته است. در چند سال اخیر استفاده از رویکردهای یادگیری عمیق به همراه تکنیک‌های داده‌کاوی رونق گرفته و همچنان در حال افزایش است.

در این بخش ابتدا به مرور روش‌های یادگیری ماشینی و داده‌کاوی در بخش تشخیص تقلب در تراکنش‌های مالی می‌پردازیم. در این بین به دلیل کاربرد بسیار زیاد روش‌های بدون نظارت (به دلیل نبودن داده برچسب‌دار)، الگوریتم‌های یادگیری بدون نظارت (خوشه‌بندی) را مورد مطالعه قرار می‌دهیم و سپس تحقیقاتی که از روش‌های یادگیری عمیق استفاده کرده‌اند را بررسی می‌کنیم. به دلیل توسعه زیاد شبکه‌های اجتماعی و ارتباط و وابستگی داده‌ها به هم و در قالب گراف و اهمیت بالای روش‌های تشخیص ناهنجاری در گراف، تحقیقات موجود در این زمینه را نیز مورد بحث و بررسی می‌گذاریم و تحقیقاتی که از این روش‌های استفاده کردند را ارزیابی خواهیم کرد.

۲-۲ روش‌های یادگیری ماشین در تشخیص تقلب

ابتدا به بررسی روش‌هایی که از داده‌کاوی و یادگیری ماشین برای تشخیص داده‌های مشکوک و رفتار مخربانه استفاده شده است می‌پردازیم و رویکردهای مورد نظر را بررسی می‌کنیم.

۲-۲-۱ ماشین بردار پشتیبان

ماشین بردار پشتیبان[1] یک روش طبقه‌بندی بانظارت است که در طبقه‌بندی خطی با استفاده ایجاد یک اَبَرصفحه[2] به تصمیم‌گیری در فضای داده‌ها می‌پردازد. عملکرد خوب در فضای داده با ابعاد بالا و زمانی که ابعاد داده‌ها از تعداد داده‌ها بیشتر است از ویژگی‌های مثبت ماشین بردار پشتیبان است. زو و همکارانش تحقیقی را جهت شناسایی تراکنش‌های متقلبانه در تراکنش‌های برخط کارت‌های اعتباری با استفاده از مدل SVM بهینه شده ارائه دادند. مدل ارائه شده از ماشین بردار پشتیبان غیرخطی و کرنل تابع پایه شعاعی برای تراکنش‌های با داده‌های خلوت استفاده می‌کند و برای شناسایی و تعیین پارامترها از الگوریتم گرید استفاده کردند. مدل ارائه شده با روش تلفیقی ID3 به همراه الگوریتم پس‌انتشار خطا به عقب مقایسه شد و عملکرد بهتری نسبت به این دو روش داشت[۱۴].

در تحقیقی با استفاده از مقایسه با سه روش نزدیک‌ترین همسایه، قانون بیز و روش ماشین بردار پشتیبان و ارزیابی به وسیله طبقه‌بندی‌کننده‌ی تجمعی مبتنی بر درخت تصمیم بر روی مجموعه داده‌های دنیای واقعی که شامل ۱۰۰۰۰۰ رکورد و ۲۰ ویژگی بودند و داده‌ها به صورت برچسب‌دار بودند، محققین دریافتند که این الگوریتم علاوه بر اینکه در مدت زمان کمتری به حالت پایدار ارزیابی می‌رسد، بلکه کارایی بسیار بالایی نیز دارد. ایده‌ی اصلی این روش استفاده از یادگیری تجمعی مبتنی بر درخت تصمیم جهت حل مشکل عدم تعادل بین داده‌های معمولی و داده‌های ناهنجار بوده است[۱۵]. در سال ۲۰۱۹ در تحقیقی، نویسندگان به ارائه روش یادگیری بانظارت با استفاده از ترکیب روش‌های ماشین بردار پشتیبان و رگرسیون خطی و رگرسیون دودویی برای افزایش دقت نرخ شناسایی، به شناسایی رفتار قانونی و مشکوک در تراکنش‌های کارت اعتباری مشتریان پرداختند. در نتایج خود نشان دادند که استفاده ترکیبی از این روش‌ها در شناسایی رفتار مشکوک عملکرد مناسبی نسبت به روش شبکه‌های پس‌انتشار داشته است و دقت آموزش در این تحقیق به ۸۰ درصد رسیده است[۱۶].

[1] Support Vector Machine (SVM)
[2] Hyperplane

دِنگ و همکارانش مدلی را برای شناسایی صورت‌حساب‌های متقلبانه با استفاده از روش ماشین بردار پشتیبان ارائه دادند. به دلیل انحراف و نامتوازن بودن داده‌ها، نویسندگان در این مقاله ابتدا به نرمال‌سازی مجموعه داده اقدام کردند تا اطمینان حاصل کنند که عدم تعادل داده بر روی عملکرد و ابعاد داده‌های ورودی تأثیری نمی‌گذارد. بعد از انتخاب ویژگی داده‌ها و کاهش ابعاد داده‌ها، به آموزش با ماشین بردار پشتیبان پرداختند. دادهای آموزشی این تحقیق شامل ۴۴ صورت حساب گزارش شده با رفتار معمولی و ۴۴ صورت حساب با برچسب متقلب بوده و برای مرحله آزمایش از ۷۳ اطلاعات صورتحساب نرمال و ۹۹ صورتحساب متقلبانه از شرکت‌های کشور چین بوده استفاده کردند. نتایج مدل ثابت می‌کند که این رویکرد نتایج قابل قبولی داشته است[۱۷].

۲-۲-۲ سامانه‌های مبتنی بر منطق فازی

منطق فازی[1] برای مدیریت نمایش ابهام و عدم قطعیت کامل محیط داده‌ها استفاده می‌شود. منطق فازی نشان‌دهنده تخمینی بودن و عدم درستی کامل حقیقت و داده‌هاست. در منطق فازی محاسبات بر پایه‌ی درجه‌ای از حقیقت صورت می‌گیرد. تصمیم‌گیری در فضای ابهام و غیرقطعی داده‌ها و تنظیم و کنترل خروجی ماشین مبتنی بر ورودی‌های چندگانه از ویژگی‌های مثبت منطق فازی است.

در سال ۲۰۱۵ تحقیقی بر روی شناسایی تقلب در تراکنش‌های کارت اعتباری با استفاده از دو رویکرد خوشه‌بندی C – Mean فازی و شبکه‌های عصبی انجام شده است. نویسندگان در این تحقیق به دلیل عدم وجود مجموعه داده تراکنش کارت اعتباری معتبر، به ساخت داده مصنوعی روی آوردند و ارزیابی‌های خود را بر روی داده‌های ساختگی خود انجام دادند. بعد از انجام خوشه‌بندی و اختصاص امتیاز به تراکنش‌های مشکوک سعی در شناسایی تراکنش‌های با الگوی متفاوت کردند. ادعا شده است که ترکیب تکنیک‌های خوشه‌بندی و مکانیسم‌های یادگیری ماشین می‌تواند بسیار در شناسایی رفتار مخربانه و کاهش ایجاد خطاهای اشتباه کمک کنید. دقت حاصل از تحقیق ۹۳,۹۰ درصد True Positive و ۶,۱

[1] Fuzzy Logic

درصد در False Positive بوده است[۱۸].

در تحقیقی [۱۹] یک مدل تلفیقی مبتنی بر روش فازی و مدل رفتار Fogg برای شناسایی رفتار متقلبانه در تراکنش‌های کارت اعتباری استفاده شده است. در این تحقیق یک سیستم فازی برای نظارت بر تاریخ فعالیت تراکنش‌ها ایجاده شده، در حالی که مدل رفتاری Fogg برای شناسایی انگیزه و توانایی تقلب افراد اعمال شده است. این مدل که FZZGY نام دارد به محاسبه درجه مشکوک بودن تراکنش‌های دریافتی افراد می‌پردازد. به دلیل ابهام و عدم قطعیت در شناسایی رفتار مشکوک داده‌ها، قوانین فازی بر روی داده‌ها اعمال شده است. نتایج حاکی از عملکرد مناسب و قابل قبول در شناسایی رفتار متقلبانه افراد است.

۲-۲-۳ مدل مارکوف پنهان

مدل مخفی مارکوف[1] یک فرآیند تصادفی دوگانه است که در مقایسه با مدل مارکوف جهت ایجاد پیچیدگی فرآیندهای تصادفی بیشتر مورد استفاده قرار می‌گیرد. مدل مخفی مارکوف یک پایه آماری بسیار قوی‌ای را داراست و می‌توان به کمک آن طول متغیرهای ورودی را مدیریت کرد و در کاربردهای مختلفی از این روش استفاده می‌شود.

آگراوال و همکارانش مدلی را برای شناسایی تقلب در تراکنش‌های کارت اعتباری مبتنی بر ترکیب روش‌های مدل مخفی مارکوف، الگوریتم ژنتیک و روش مبتنی بر رفتار ارائه دادند. مدل ارائه شده از سه گام تشکیل شده است. در ابتدا با استفاده از مدل مخفی مارکوف گزارشی از تراکنش‌های قبلی را نگهداری کردند، سپس با کمک تکنیک مبتنی بر رفتار به خوشه‌بندی پروفایل کاربران و تراکنش‌های آن‌ها پرداختند و در سه دسته‌ی کم، متوسط و زیاد خوشه‌بندی کردند. و در نهایت با الگوریتم ژنتیک به محاسبه مقدار حد آستانه پرداختند[۲۰].

در تحقیقی [۲۱] نویسندگان با استفاده از مدل مخفی مارکوف به بررسی رفتار دارندگان کارت

[1] Hidden Markov Model

اعتباری و تراکنش‌های آن‌ها پرداختند. در این تحقیق با روش مدل مارکوف مخفی، تعداد تراکنش‌ها بر حسب مقدار آن‌ها به سه دسته کم، متوسط و زیاد تقسیم گردید، سپس با خوشه‌بندی و ایجاد یک حد آستانه به تمیز دادن بین الگوی داده‌های قانونی و متقلبانه پرداختند. محققین در این تحقیق پی بردند که با استفاده از روش مدل مخفی مارکوف می‌توان پیچیدگی‌های سامانه تشخیص تقلب را کاهش داد. علاوه بر این نیز استفاده از روش مدل مخفی مارکوف توانایی مدیریت و اجرا بر روی حجم زیاد داده‌ها را دارد.

لیر و همکارانش در تحقیق خود رویکردی را ارائه کردند که موجب افزایش دقت و کارایی در شناسایی تراکنش‌های متقلبانه در کارت اعتباری می‌شود. رویکرد ارائه شده با استفاده از مدل مخفی مارکوف و تکنیک خوشه‌بندی K-Means جهت پیدا کردن نزدیک‌ترین مرکز خوشه‌ها و ترکیب آن‌ها در یک خوشه می‌گردد. مدل مخفی مارکوف با استفاده از الگوریتم باوم وِلچ[1] به آموزش داده‌ها می‌پردازد. نتایج حاصل از الگوریتم اشتباهات نسبتاً کمی را در بین داده‌ها برای تشخیص داده‌های متقلبانه ایجاد می‌کند[۲۲].

۲-۲-۴ الگوریتم نزدیک‌ترین همسایه

تکنیک نزدیک‌ترین همسایه[2] یک روش غیر پارامتری برای طبقه‌بندی و رگرسیون داده‌ها است. هدف در این روش پیدا کردن نزدیک‌ترین همسایه‌ها مبتنی بر شباهت مجموعه داده‌ها و تولید نقطه جدید نمونه مبتنی بر معیار اندازه فاصله بین دو نمونه داده است. از معروف‌ترین معیارهای محاسبه فاصله، معیار فاصله اقلیدسی[3] است. معیار فاصله همینگ، معیار فاصله منهتن و معیار فاصله Minkowski از دیگر معیارها در این الگوریتم است. سادگی پیاده‌سازی این الگوریتم از مزایای روش نزدیک‌ترین همسایه است.

در تحقیقی [۲۳] نویسندگان با ترکیب روش نزدیک‌ترین همسایه و تشخیص تعامل خودکار توزیع مربع خی‌دو مدلی را برای شناسایی رفتار متقلبانه در کارت‌های اعتباری ارائه کردند. محققین کارایی سه

[1] Baum Welch algorithm
[2] Nearest neighbor algorithm
[3] Euclidean Distance

روش یادگیری ماشین که بر روی داده‌های کارت اعتباری برای تشخیص تقلب ارائه شده را مورد بررسی قرار دادند. یک تحلیل مقایسه‌ای با استفاده از رگرسیون خطی، روش بیز ساده و روش نزدیک‌ترین همسایه روی ۲۸۴,۸۴۷ تراکنش دارندگان کارت اعتباری اروپایی انجام شده و نتایج حاصل از آن نشان‌دهنده‌ی دقت بسیار بالای روش نزدیک‌ترین همسایه با ۹۷/۶۹ درصد نسبت به دیگر تکنیک‌ها بوده است[۲۴].

۲-۲-۵ شبکه بیزین

شبکه بیزین[1] یکی از مدل‌های گرافیکی‌ای است که با استفاده از وابستگی‌های شرطی و روابط مستقل بین متغیرهای مختلف عمل می‌کند. مفهوم اصلی شبکه‌های بیزین مانند گره‌ها و یال در گراف جهت‌دار است. شبکه‌های بیزینی در شناسایی محاسبات احتمالاتی بخصوص در حالتی که غیرقطعیت وجود دارد کاربرد دارد.

در تحقیقی [۱۷] محققین با بکارگیری طبقه‌بندی کننده بیز ساده به طراحی مدلی جهت شناسایی صورت‌حساب‌های مالی متقلبانه پرداختند. هدف این تحقیق افزایش دقت طبقه‌بندی بوده است. نتایج حاصل از ارزیابی‌ها که بر روی مجموعه دادگانی با ۴۴ داده با رفتار معمولی و متقلبانه بوده، حاکی از دقت بالا در شناسایی رفتار افراد متقلب در دادگان بود. هاجک و همکارانش با به دست آوردن ویژگی‌های خاص در گزارش‌های صورت‌حساب‌های مالی در شرکت‌ها و اعمال ۱۴ روش یادگیری ماشین دریافتند که روش شبکه‌های بیزین از دیگر رویکردهای عملکرد بهتری نشان داده است[۲۵].

۲-۲-۶ شبکه‌های عصبی مصنوعی

شبکه عصبی مصنوعی[2] مشابه با مغز انسان با استفاده از روش‌های غیرخطی به شناسایی و دسته‌بندی داده‌ها می‌پردازد. جزء اصلی شبکه‌های عصبی نورون‌ها هستند که ساختار درونی لایه‌های محاسباتی را

[1] Baysian Network
[2] Artificial Nueral Network

تشکیل می‌دهند.

در تحقیقی با استفاده از شبکه‌های عصبی به شناسایی تقلب و رفتار و الگوی مخربانه در معاملات و تراکنش‌ها پرداختند. شبکه عصبی مصنوعی استفاده شده در این تحقیق شبکه عصبی خودسازمانده[1] بوده است. [۲۶]. قبادی و همکارانش با استفاده مدلی تلفیقی به شناسایی تقلب در تراکنش‌های مالی با استفاده از رویکردهای شبکه عصبی پرداختند. مدل ارائه شده را بر روی مجموعه دادگان دنیای واقعی ارزیابی و آزمایش نمودند. مدل ارائه شده با داده‌های نامتوازن عملکرد خوبی نشان نداد، از این رو محققین از روش دیگری برای حل این مشکل پرداختند. نتایج حاصل از تحقیق نشان‌دهنده‌ی افزایش نرخ دقت و کاهش هزینه‌های اشتباه‌های منفی شده است [۲۷].

ساهین و همکاران در تحقیقی با به کار گرفتن روش شبکه عصبی مصنوعی و رگرسیون خطی و اعمال بر روی مجموعه دادگان دنیای واقعی دریافتند که روش شبکه‌های عصبی مصنوعی با افزایش دقت بالایی روبه‌رو هستند و از دیگر روش‌های عملکرد بهتری دارند [۲۸].

جدول ۰-۱:رویکردهای مختلف تشخیص ناهنجاری و فرضیات آن‌ها

روش تشخیص ناهنجاری	فرضیات	امتیاز ناهنجاری داده	مثال‌های قابل‌توجه
رویکردهای آماری	با داشتن یک مدل تصادفی، داده‌های نرمال در مناطق با احتمال بالا قرار دارند و داده‌های ناهنجار در مناطق با احتمال کمتر	با توجه به توزیع احتمالاتی داده‌هایی که در مناطق با احتمال بیشتری قرار دارند امتیاز کمتری دارند.	مدل رگرسیون و شبکه بیز و مارکوف پنهان
طبقه‌بندی	*وجود داده‌های برچسب دار *یادگیری طبقه‌بندی و جداسازی داده‌های نرمال از داده‌های ناهنجار	معیار طبقه‌بند جهت تخمین اینکه داده‌ها به کدام کلاس تعلق دارند	ماشین بردار پشتیبان تک‌کلاسه، شبکه عصبی
نزدیک‌ترین همسایه	داده‌های نرمال در همسایگی‌های با چگالی بیشتر اتفاق می‌افتد و داده ناهنجار در نقاط دورتری	فاصله از k-امین نزدیک‌ترین همسایه	K-نزدیک‌ترین همسایه

[1] Self Organizing Map

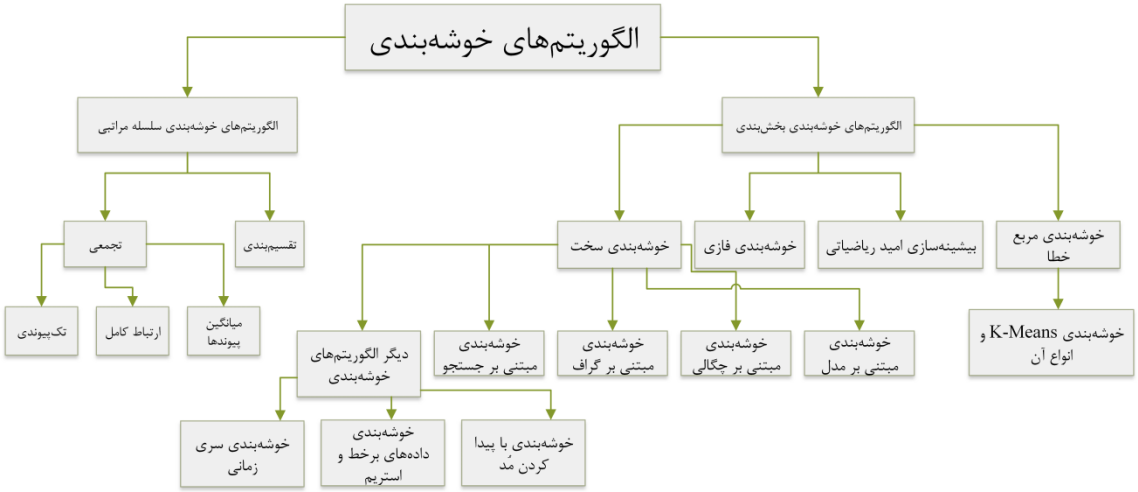
	هستند		
خوشه‌بندی	داده‌های نرمال به خوشه‌های مختلف تعلق دارند، در صورتی که داده‌های ناهنجار به هیچ خوشه‌ای تعلق ندارند	فاصله از مرکز نزدیک‌ترین خوشه	بیشینه‌سازی امید (Expectation Maximization)، روش‌های خوشه‌بندی مانند K-Means و نقشه خودسازماندهی‌شده (SOM)

۲-۳ روش‌های مبتنی بر خوشه‌بندی

همانطور که در فصل اول صحبت شد، به دلیل نبودن داده برچسب‌دار، استفاده از الگوریتم‌های خوشه‌بندی[1] و یادگیری بدون نظارت داده‌ها در حوزه تشخیص ناهنجاری و به خصوص در حوزه تقلب در تراکنش‌های مالی از اهمیت زیادی برخوردار است و تحقیقات زیادی در این زمینه صورت گرفته و محققین همچنان در حال استفاده از تکنیک‌های خوشه‌بندی برای شناسایی رفتار مشکوک در تراکنش‌ها هستند.

یادگیری بدون نظارت تلاش می‌کند تا با توجه به ساختارهای پنهان بین داده‌های بدون برچسب، آن‌ها را در خوشه‌های مختلف جای دهد. در تکنیک‌های خوشه‌بندی برای پیدا کردن و استخراج قوانین خوشه‌ها و شباهت بین آن‌ها، نیاز به داده‌ای که از قبل برچسب داشته باشد نیست. روش‌های مختلفی برای خوشه‌بندی داده‌ها وجود دارد که در شکل ۲-۱ مشاهده می‌کنیم.

[1] Clustering Technique



شکل ۰-۱: الگوریتم‌های خوشه‌بندی

۲-۳-۱ مراحل خوشه‌بندی

انتخاب و یا استخراج ویژگی: اولین مرحله بعد از ورود داده‌ها و قبل از خوشه‌بندی مرحله انتخاب و یا استخراج ویژگی است. انتخاب ویژگی به معنی انتخاب مجموعه‌ای از ویژگی‌ها از کل مجموعه دادگان، در حالی که استخراج ویژگی به معنی تبدیل و یا انتقال داده‌ها جهت تولید ویژگی‌های جدید از دادگان اصلی. انتخاب و یا استخراج ویژگی مرحله بسیار مهم در خوشه‌بندی مؤثر و کارا است. انتخاب درست ویژگی‌ها (که مصون از نویز باشند و به راحتی تفسیر بشوند) از اهمیت زیادی برخودار است و می‌تواند باعث کاهش پیچیدگی محاسباتی و ساده‌سازی فرآیند بعدی بشود.

الگوریتم خوشه‌بندی: برای تعریف هر الگوریتم مقدماتی وجود دارد. با هر مجموعه داده $X = x_1, x_2, \ldots, x_N$ خوشه‌بندی در جهت پارتیشن‌بندی مجموعه داده X است به صورتی که $C = C_1, C_2, \ldots, C_k (K \leq N)$ باشد و شرایط زیر برقرار باشد:

- $C_i \neq \emptyset, i = 1, \ldots k$ بدین معنی که یک خوشه نمی‌تواند خالی باشد و حتماً باید حداقل یک داده در آن وجود داشته باشد.
- $\bigcup_{i=1}^{k} C_i = X$: همه‌ی داده‌ها باید خوشه‌بندی بشوند و متعلق به خوشه‌ای باشند.
- $C_i \cap C_j = \emptyset, i, j = 1, \ldots, K \; and \; i \neq j$:هر داده باید متعلق به فقط و فقط یک خوشه

باشد. اگرچه روش‌های دیگر خوشه‌بندی مانند خوشه‌بندی فازی وجود که یک داده می‌تواند در بیشتر از یک خوشه عضو باشد.

- $d(x_i, x_j)$ در C_i باید حداقل باشد؛ بدین معنی که فاصله بین داده‌ها درون یک خوشه باید در کمترین حالت خود باشد.
- $d(C_i, C_j)$ باید در بیشترین حالت خود باشد؛ بدین معنی که فاصله بین خوشه‌ها باید حداکثر باشد.

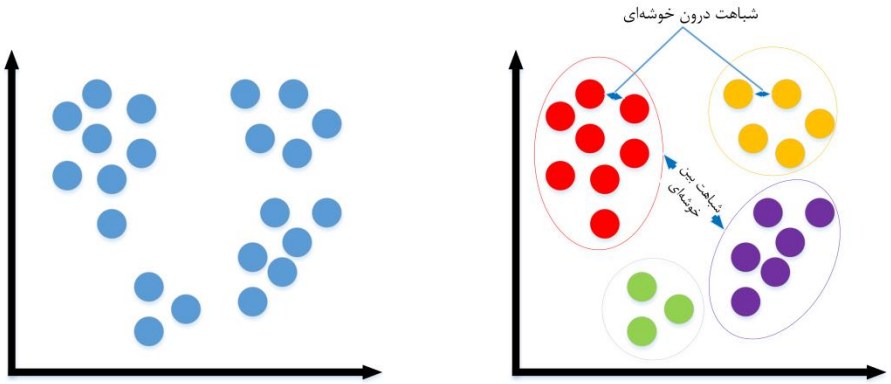
انتخاب الگوریتم خوشه‌بندی: یکی از مهم‌ترین مراحل، انتخاب نوع الگوریتم خوشه‌بندی با توجه به داده‌های ورودی مد نظر است. معیار شباهت و یا عدم شباهت که همجواری نیز نامیده می‌شود، به انتخاب الگوریتم خوشه‌بندی با کیفیت بالا نیز کمک می‌کند. انتخاب الگوریتم خوشه‌بندی مناسب منجر به بخش‌بندی درست داده‌ها و دقت بالای الگوریتم‌های خوشه‌بندی می‌شود.

جدول ۰-۲: مقایسه الگوریتم‌های خوشه‌بندی، n تعداد داده‌ها، k تعداد خوشه، d ابعاد و m پارامتر مربوط به شبکه‌های خودسازمانده[۲۹]

الگوریتم خوشه‌بندی	$K-Means$	سلسله مراتبی	بیشینه‌ساز ی امید	BIRCH	CURE	مبتنی بر چگالی	شبکه‌ها خودسازمانده
عملکرد با تعداد داده زیاد	خوب	متوسط	خوب	خوب	خوب	خوب	متوسط
پیچیدگی زمانی	$O(knd)$	$O(n^2)$	$O(knd)$	$O(n)$	$O(n^2 \log n)$	$O(n^2)$	$O(n^2 m)$
دقت	کم	زیاد	کم	زیاد	زیاد	کم	زیاد
توانایی مدیریت داده‌های	خیر	بله	خیر	بله	بله	خیر	خیر

نویز

اعتبار و ارزیابی خوشه‌ها: الگوریتم‌های مختلف خوشه‌بندی، خوشه‌های مختلفی را ایجاد می‌کنند؛ بنابراین ضروری است که کیفیت خوشه‌های ایجاد شده مورد ارزیابی قرار گیرد. برای ارزیابی درست داده‌ها در خوشه‌های مختلف سه معیار نسبت، درونی، خارجی وجود دارد و براساس دانش قبلی از ساختار خوشه‌ها و یا بدون دانش قبلی نمایش درست خوشه‌ها را انتخاب می‌کنند.



شکل ۲-۰:نمای کلی خوشه‌بندی داده‌ها و شباهت بین خوشه و میان خوشه‌ای

تفسیر نتایج: هدف خوشه‌بندی ایجاد اطلاعات معنی‌دار از داده‌ها برای کاربران است. بدین معنی که فرد متخصص و تحلیلگر داده‌ها با آنالیز و تفسیر خوشه‌ها به تصمیم‌گیری و راه‌حل در مورد مسئله مورد نظر می‌پردازد.

۲-۳-۲ فرضیات کلیدی در شناسایی داده‌های ناهنجار با استفاده از خوشه‌بندی

همانطور که پیش‌تر گفته شد، هدف خوشه‌بندی بخش‌بندی و در یک گروه قرار دادن داده‌ها مشابه و شبیه به هم هست، و این تکنیک می‌تواند در شناسایی الگوی داده ناهنجار به ما در مجموعه دادگان کمک کند. در همین راستا سه فرضیه اساسی وجود دارد که باید توجه بشوند:

فرضیه اول: با الگوریتم‌های خوشه‌بندی ما فقط داده‌هایی که نرمال هستند و الگوی رفتار مخرب

ندارند را می‌توانیم خوشه‌بندی کنیم و داده‌هایی که به این خوشه‌ها تعلق نداشته باشند به عنوان داده ناهنجار در نظر گرفته می‌شوند. برای مثال خوشه‌بندی مبتنی بر چگالی، داده‌های نویز را در هیچ خوشه‌ای جای نمی‌دهد و نویزها به صورت داده ناهنجار در نظر گرفته می‌شوند[۳۰].

فرضیه دوم: بعد از اتمام مراحل خوشه‌بندی اگر خوشه‌ای شامل داده‌های نرمال و داده ناهنجار باشد، فرض بر این است که داده‌های نرمال به مرکز خوشه نزدیک‌تر هستند و داده‌هایی که الگوی متفاوت دارند و ناهنجار محسوب می‌شود از مرکز داده دورتر هستند[۳۱]. به عنوان مثال در [۳۲] Svetlona و همکارانش الگوریتمی مبتنی بر خوشه‌بندی برای حذف داده‌های پرت را ارائه دادند که به صورت همزمان به خوشه‌بندی داده‌ها و شناسایی داده پرت می‌پرداخت. الگوریتم ارائه شده آن‌ها دو مرحله داشت؛ در مرحله اول اعمال الگوریتم خوشه‌بندی $K-Means$ و سپس معیار و امتیاز پرت بودن داده را اندازه می‌گرفتند. اگر معیار پرت بودن داده از مقدار تعیین شده حد آستانه بالاتر بود، آن داده به عنوان داده ناهنجار در نظر گرفته می‌شد و از مجموعه دادگان حذف می‌گردید. مجموعه دادگان محققین در این تحقیق داده‌های ساختگی و برخی عکس‌های نقشه بود. و معیار ارزیابی میانگین خطای مطلق برای ارزیابی کارایی الگوریتم را استفاده کردند.

فرضیه سوم: در خوشه‌بندی، خوشه‌ها دارای اندازه‌های مختلفی هستند؛ خوشه‌های کوچک‌تر و تُنُک‌تر به عنوان ناهنجاری و خوشه‌های بزرگتر و با داده بیشتر به عنوان داده نرمال هستند. نمونه‌هایی که به خوشه‌هایی که با سایز و یا چگالی کمتر از حد آستانه‌ی تعیین شده قرار دارند به عنوان ناهنجاری هستند. نویسندگان [۳۳] تعریفی را برای ناهنجاری‌های محلی مبتنی بر خوشه‌بندی ارائه دادند. آن‌ها از پارامترهای عددی برای شناسایی خوشه‌های کوچک و خوشه‌های بزرگ استفاده کردند. از الگوریتم SQUEEZER برای خوشه‌بندی داده‌ها استفاده کردند که خوشه‌های با کیفیت بالا تولید می‌کند و توانایی مدیریت داده‌ها با ابعاد بالا را نیز دارد. سپس از الگوریتم CBLOF برای شناسایی معیار پرت بودن هر رکورد داده استفاده کردند.

۲-۳-۳ مروری بر پژوهش‌های مبتنی بر خوشه‌بندی

آیزا و همکاران روش تشخیص ناهنجاری‌ای را برای شناسایی بازپرداخت‌های متقلبانه توسط الگوریتم خوشه‌بندی $K-Means$ ارائه کردند. مجموعه دادگان مورد استفاده تراکنش‌های بازپرداخت شرکت‌های ارتباط از راه دور بود و نیاز به پیش‌پردازش داده‌ها برای بهبود الگوریتم بود. با مقدار پارامترهای مختلف برای تعداد خوشه‌ها در نرم‌افزار WEKA به نتایج حاصل از خوشه‌ها و تحلیل آن‌ها پرداختند و داده‌های مشکوک را شناسایی کردند[۳۴].

در سال ۲۰۱۱ محققین رویکردی را برای شناسایی تقلب در ارزیابی ادعاهای گروه‌های بیمه‌ای انجام دادند. ایده اصلی رویکرد مانند فرضیه سوم در بخش قبل بوده و ویژگی بارز خوشه‌های مشکوک، حجم زیاد پرداختی‌های خیرخواهانه، تعداد بالای آن‌ها و فرآیند زمانی طولانی آن‌ها بوده است. مجموعه دادگان شامل ۴۰۰۸۰ ادعاهای بیمه بوده که در سال ۲۰۰۹ پرداخت شده است. الگوریتم $K-Means$ توسط WEKA و نرم‌افزار تحلیل داده SPSS نیز اجرا شد و خوشه‌های با تعداد کمتر به عنوان تقلب‌های احتمالاتی در نظر گرفته‌شده‌اند. تحلیل اینکه خوشه‌های انتخاب شده به درستی نشان‌دهنده‌ی رفتار مشکوک داده‌ها باشند توسط تحلیلگران داده انجام شده است[۳۵].

نویسندگان در مقاله [۳۶] استفاده از الگوریتم $X-Means$ (که مزیت آن نسبت به الگوریتم $K-Means$ شناسایی خودکار تعداد خوشه‌ها است) را برای شناسایی الگوی متفاوت داده‌ها مورد مطالعه قرار دادند. الگوریتم $X-Means$ از معیار اطلاعات بیزین[1] برای شناسایی تعداد خوشه‌ها استفاده می‌کند. در این تحقیق نیز از نرم‌افزار WEKA [۳۷] برای خوشه‌بندی و شناسایی رفتار متقلبین در مزایده برخط در شرکت یاهو استفاده شد.

یک مورد مطالعه بر روی شناسایی پولشویی با استفاده از ترکیب داده‌کاوی و تکنیک‌های محاسباتی توسط Njien و همکاران صورت گرفته است. به دلیل سادگی روش خوشه‌بندی از این روش استفاده شده و تعداد خوشه‌ها توسط متخصصین حوزه تشخیص پولشویی به صورت یک عدد ثابت در نظر گرفته شد.

[1] Bayesian Information Criterion (BIC)

مجموعه داده‌های این تحقیق شامل میلیون‌ها تراکنش شش شرکت سرمایه‌گذاری که اطلاعات مشتریان به صورت محرمانه بود و نتایج قابل قبولی را در شناسایی الگوی متفاوت داده‌ها داشته است[۳۸].

نویسندگان در تحقیقی [۳۹] رویکردی را برای شناسایی سه مورد تقلب در حوزه خرید ارائه دادند. پرداخت چندباره‌ی صورتحساب، تغییر دستور خرید و انحراف دستور خرید. مجموعه دادگان مورد استفاده در این تحقیق از سامانه‌های برنامه‌ریزی منابع سازمانی جمع‌آوری‌شده و شامل اطلاعاتی در مورد دستورات خرید، فاکتور کالا و صورتحساب‌ها بوده است. با استفاده از الگوریتم خوشه‌بندی $K-Means$ و تحلیل نتایج آن دریافتند که نرخ میانگین بالای برخی از ویژگی‌های خاص به همراه خوشه‌های کوچک‌تر به عنوان داده مشکوک و ناهنجار در نظر گرفته می‌شوند.

در [۴۰] نویسندگان از الگوریتم خوشه‌بندی مبتنی بر چگالی[1] برای شناسایی تقلب در تراکنش‌های کارت اعتباری استفاده کردند. این الگوریتم خوشه‌هایی را از داده‌ها ایجاد می‌کند و می‌تواند داده نویز را از داده اصلی جدا کند. در این تحقیق داده نویز به عنوان رفتار مشکوک و متقلبانه در کارت اعتباری در نظر گرفته شد. فرضیه موجود در این مورد مطالعاتی بدین صورت بوده است که اگر داده‌ای به خوشه‌ای تعلق نداشت به عنوان داده ناهنجار شناخته می‌شود. سپس با تعریف درجه‌ی تقلب و تعریف یک حد آستانه به شناسایی داده‌ها و خوشه‌ها ناهنجار پرداختند. مجموعه داده‌ی این تحقیق داده‌های ساختگی از تاریخچه‌ی تراکنش‌ها و رفتار درست دارندگان کارت اعتباری بوده است.

استفاده از الگوریتم خوشه‌بندی متراکم سلسله مراتبی برای تشخیص تقلب در تراکنش‌های مالی توسط Togo و همکاران ارائه شده است. فرضیه موجود در این تحقیق بدین صورت است که داده با رفتار و فعالیت مخربانه در یک خوشه با داده‌های عادی قرار نمی‌گیرد. با استفاده از اختصاص امتیاز به داده پرت، به محاسبه میزان فعالیت مخربانه و پرت بودن داده‌ها پرداختند و سپس خوشه‌بندی داده‌ها انجام شد. مجموعه داده مورد استفاده این تحقیق از موسسه بین‌المللی آمار گرفته شده که شامل تراکنش‌های مالی شرکت‌های خارجی دیگر کشورها بوده است. و شامل ۴۱۰۰۰۰ تراکنش گزارش شده در هشت ماه

[1] Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

بود [۴۱].

۲-۴ روش‌های مبتنی بر یادگیری عمیق

استفاده از رویکردهای یادگیری عمیق در تشخیص ناهنجاری در سال‌های اخیر بین محققین بسیاری، از محبوبیت بالایی برخوردار گشته است. استفاده از یادگیری عمیق نتایج بسیار بهتری را نسبت به روش‌های پیشین یادگیری ماشینی و داده‌کاوی می‌دهد و کاربردهای نسبتاً زیادی در حوزه‌ها مختلف دارد. در زیر برخی از دلایل استفاده از روش‌های مبتنی بر یادگیری عمیق را در شناسایی ناهنجاری ذکر کرده‌ایم:

- کارایی بالا
- ساخت مدل‌های پیچیده و غیرخطی
- کار با تعداد زیاد داده‌ها
- کار با داده‌های چند متغیره
- مدیریت و استخراج اطلاعات از داده‌های با ابعاد بالا
- کمترین نیاز به تنظیم پارامترها برای رسیدن به نتایج مناسب

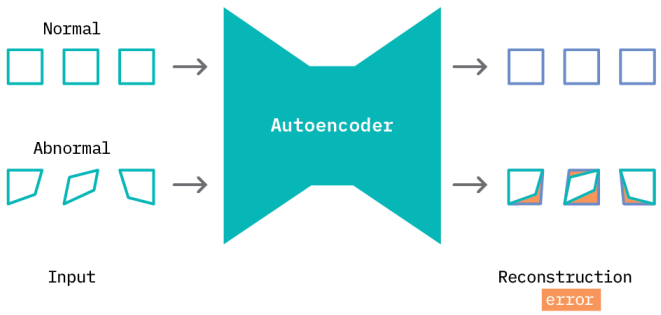
در این بخش به یک مرور کلی بر مدل‌های معماری یادگیری عمیق می‌پردازیم و اینکه چگونه این معماری‌ها در تشخیص ناهنجاری استفاده می‌شوند. درواقع اکثر روش‌ها به یادگیری مدلی برای شناسایی داده‌ها نرمال می‌پردازند و سپس برای شناسایی فعالیت‌ها مخربانه به امتیازبندی داده‌ها اقدام می‌کنند تا داده ناهنجار را شناسایی کنند.

بیشتر رویکردهای یادگیری عمیق در حوزه مدل‌های رمزگذار-رمزگشا[1] قرار می‌گیرد. یادگیری بازنمایی و بردار ویژگی‌های آن‌ها از داده‌های ورودی توسط رمزگذار و سپس تلاش برای بازسازی دوباره داده‌های ورودی اصلی با توجه به نمایش داخلی داده‌های توسط رمزگشا رویکرد این‌گونه روش‌هاست.

[1] Encoder-Decoder Models

۲-۴-۱ شبکه‌های خودرمزگذار

شبکه‌های خودرمزگذار[1] برای یادگیری بازنمایی[2] از داده‌های ورودی اصلی طراحی شده‌اند. این شبکه‌ها از دو جزء اصلی رمزگذار و رمزگشا تشکیل شده است. رمزگذار برای یادگیری نمایش با ابعاد پایین داده، و رمزگشا برای یادگیری و تبدیل بازنمایی حاصل از رمزگذار به داده‌های اصلی ورودی ایجاد شده است. هدف اصلی این نوع شبکه‌ها، کمینه کردن خطای بازسازی داده‌ها توسط رمزگشا است که تفاوت بین داده ورودی و داده بازسازی شده است و توسط معیار میانگین مربع خطا[3] اندازه‌گیری می‌شود. از کاربردهای این نوع شبکه: کاهش ابعاد، کاهش نویز از تصاویر، استخراج ویژگی بدون ناظر و فشرده‌سازی داده استفاده می‌شود. در شکل ۲-۳ معماری این نوع شبکه‌ها را مشاهده می‌کنیم.



شکل ۰-۳: معماری شبکه خودرمزگذار[۴۲]

در پژوهشی که در سال ۲۰۱۶ توسط Silivio L. Domingos و همکارانش بر روی داده‌های سیستم خریدوفروش دولتی کشور برزیل صورت گرفته است، با استفاده از ابزار داده‌کاوی و یادگیری ماشین H2O و مدل‌های یادگیری عمیق تعبیه‌شده در این ابزار به بررسی ناهنجاری‌های موجود در خریدهای دولتی پرداخته‌اند. مدل یادگیری عمیق استفاده شده در این ابزار، شبکه خودرمزگذار بوده است که با تغییر پارامترهای آن به ارزیابی نتایج حاصل پرداخته‌اند و بهترین خطای بازسازی داده در شبکه خودرمزگذار دستیابی داشته‌اند. داده‌های این پژوهش، مجموعه دادگان واقعی خریدوفروش دولتی دولت برزیل در سال‌های ۲۰۱۴ و ۲۰۱۵ بوده است. نتایج حاصل از ارزیابی این روش که با تغییر پارامترهای روش

[1] Autoencoder
[2] Representation Learning
[3] Mean Square Error

یادگیری عمیق بوده در مرحله آموزش به دقت ۰,۰۰۱۲ در میانگین مربعات خطا رسیده و از این آموزش در داده‌ها تست برای شناسایی داده‌ها ناهنجار استفاده شده است[۴۳].

در سال ۲۰۱۸ Tsatsral Amarbayasgalan و همکارانش الگوریتمی را ارائه دادند که با یادگیری بدون نظارت و با استفاده از شبکه یادگیری عمیق خودرمزگذار و خوشه‌بندی مبتنی بر فاصله به تشخیص ناهنجاری می‌پردازد. الگوریتم ارائه شده آن‌ها که به طور خلاصه DAE-DBC نام دارد و در داده‌های بدون برچسب با ابعاد بالا عملکرد خوبی از خود نشان داده. یکی از مشکلات اصلی روش ماشین بردار پشتیبان تک کلاسه که برای تشخیص داده ناهنجار استفاده می‌شود، پیچیدگی زمانی و مسئله حافظه است که با افزایش ابعاد داده و تعداد داده‌ها به خوبی عمل نمی‌کند. و مشکل اصلی‌ای که روش‌های خوشه‌بندی به‌تنهایی دارند این است که کارایی این روش‌های بسیار به الگوریتم خوشه‌بندی استفاده شده دارد. برای حل این مشکل محققین بسیاری استفاده از روش‌های تحلیل مؤلفه‌ی اصلی را برای کاهش ابعاد داده ارائه کرده‌اند[۴۴].

زمینی و همکاران به شناسایی تقلب به صورت بدون ناظر با استفاده از خوشه‌بندی مبتنی بر شبکه‌های خودرمزگذار پرداختند. شبکه خودرمزگذار دارای سه لایه پنهان بوده (که در لایه اول ۲۵، در لایه دوم ۲۰ و در لایه سوم ۱۰ نورون وجود دارد) و از خوشه‌بندی $K-Means$ که با تعداد خوشه‌های ۲ و در ۳۰۰ مرحله استفاده شد. مجموعه دادگان در تحقیق شامل ۲۸۴۸۰۷ تراکنش شرکت‌های اروپایی بوده است. نتایج حاصل حاکی از دقت بالایی رویکرد ارائه‌شده با دقت ۹۸,۹ درصد است که در مقایسه با دیگر روش‌ها عملکرد بهتری داشته است [۴۵].

در تحقیقی [۴۶] نویسندگان با توجه به تنوع، حجم داده، ابعاد داده و ماهیت غیر ساختاری داده‌ها از رویکری یادگیری عمیق مبتنی بر شناسایی رفتار خوشه‌ها برای تشخیص تقلب استفاده کردند. با استفاده از روش LSTM دوگانه به تعبیه رشته و تاریخچه‌ی تراکنش‌ها در حالت بدون ناظر پرداختند و با الگوریتم خوشه‌بندی سلسله مراتبی مبتنی بر چگالی به خوشه‌بندی میلیون‌ها تراکنش با کمک پردازنده

گرافیکی پرداختند. سپس با تحلیل خوشه‌ها و شناسایی خوشه‌های ریسکی رفتار و الگوها متقلبانه در بین داده‌های خوشه‌ها بودند. داده‌های این تحقیق داده‌های دنیای واقعی بودند.

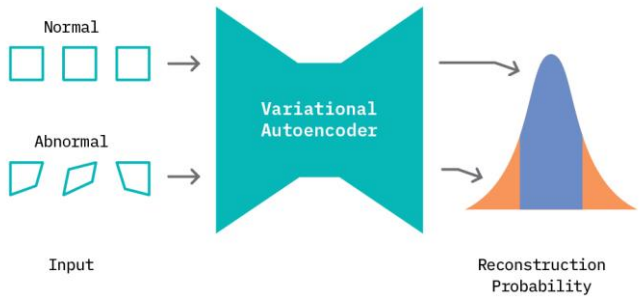
در پژوهشی در سال۲۰۲۰ نویسندگان از رویکردی دو مرحله‌ای برای شناسایی تقلب در تراکنش‌های کارت اعتباری استفاده کردند. در مرحله‌ی اول با استفاده از یک شبکه خودرمزگذار به تبدیل ویژگی‌های تراکنش به یک بردار ویژگی پرداختند. دلیل این امر حذف ویژگی‌های غیرضروری است که ممکن است باعث عملکرد ضعیف‌تر الگوریتم‌های طبقه‌بندی بشود عنوان شده است. سپس در مرحله‌ی بعد بردار ویژگی به دست آمده را به عنوان ورودی یک طبقه‌بندی کننده استفاده کردند. مجموعه دادگان مورد استفاده در این تحقیق از تراکنش‌های دارندگان کارت اعتباری اروپا در سال ۲۰۱۳ بوده است و شامل ۲۸۴۸۰۷ تراکنش است. نتایج تحقیق نشان می‌دهد که روش ارائه‌شده دارای معیارها Precision و Recall قابل قبولی نسبت به روش‌های مختلف طبقه‌بندی و متدهای مختلف شبکه‌های خودرمزگذار است [۴۷].

نویسندگان در تحقیقی [۴۸] با استفاده از یک شبکه خودرمرگذار و جنگل تصادفی احتمالاتی[1] به شناسایی تقلب در تراکنش‌های کارت‌های اعتباری پرداختند. در ابتدا با استفاده از شبکه خودرمزگذار به استخراج ویژگی‌های تراکنش‌های اعتباری پرداختند تا ابعاد داده‌های ورودی کوچک‌تر شود. سپس با جنگل تصادفی و یک مکانیسم یادگیری تجمعی[2] با استفاده از مفهوم یادگیری کیسه‌ای[3] به شناسایی داده‌های نرمال و غیر نرمال پرداختند. داده‌های این تحقیق تراکنش‌های دارندگان کارت اعتباری اروپایی هستند که به دلیل عدم تعادل بین داده‌های با برچسب متقلب و داده‌های نرمال از تکنیک‌های نمونه‌برداری برای تعادل بین داده‌ها استفاده کردند. نتایج حاصل از الگوریتم نشان‌دهنده‌ی عملکرد مناسب این رویکرد بر روی داده‌ها است و دقت و کارایی بالایی دارد.

[1] Probabilistic Random forest
[2] Ensemble Learning
[3] Bagging

۲-۴-۲ شبکه‌های خودرمزگذار متغیر[1]

این نوع شبکه‌ها به نوعی ادامه‌ای از شبکه‌های خودرمزگذار هستند و همانند آن دارای رمزگذار و رمزگشا هستند، اما دارای ساختاری متفاوتی در استنتاج متغییر هستند. این شبکه‌ها توزیعی از داده‌های ورودی را آموزش می‌بینند و در قسمت رمزگشا با کمک نمونه‌گیری از توزیع داده‌های ورودی به بازسازی داده‌ها می‌پردازد. همانند روش‌های بیزین، در این شبکه‌ها نیز مفاهیم اطلاعات پیشین، مشابهت و اطلاعات پسین کاربرد دارد و رمزگشا به یادگیری پارامترهای توزیع ورودی داده‌ها (میانگین و واریانس) می‌پردازد. ماهیت توزیع داده‌های ورودی (توزیع گاووسی توزیع برنولی) وابسته به نوع داده‌ها است. کمینه کردن خطای بین تخمین توزیع ورودی تولید شده و توزیع واقعی داده‌ها توسط معیار همگرایی Kullback-Leibler، معماری کلی این‌گونه شبکه‌هاست. این معیار فاصله بین توزیع‌ها را در نظر می‌گیرد و اینکه چه مقدار اطلاعات توسط یک توزیع از دست رفته را محاسبه می‌کند. همانند شبکه‌های خود رمزگذار این نوع شبکه‌ها نیز کاربردهایی در کاهش ابعاد، انتخاب ویژگی بدون ناظر و حذف نویز از تصاویر دارند.



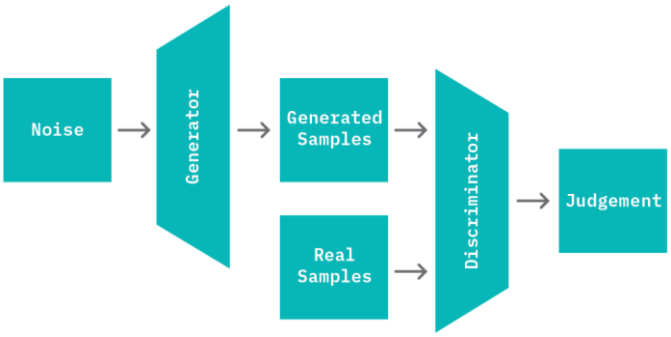
شکل ۰-۴: معماری شبکه خودرمزگذار متغیر[۴۲]

۲-۴-۳ شبکه‌های متخاصم مولد

شبکه‌های متخاصم مولد[2] شبکه‌های عصبی هستند که برای یادگیری توزیع داده‌های ورودی تولید شده استفاده می‌شوند. از دو جزء مولد و متمایزکننده تشکیل شده‌اند. در قسمت مولد با یادگیری توزیع

[1] Variational Autoencoder
[2] Generative Adversarial Networks (GAN)

داده‌های ورودی به تولید داده نزدیک به آن می‌پردازد و متمایزکننده به تشخیص داده تولید شده و مقایسه آن با داده اصلی می‌پردازد. در مرحله مولد در هر بار آموزش شبکه باید داده‌ای را تولید کرد که برای قسمت متمایز کننده قابل‌تشخیص نباشند و همانند داده اصلی باشند.



شکل ۰-۵: معماری شبکه‌های مولد متخاصم[۴۲]

همانطور که پیش‌تر گفته شد، یکی از مشکلات اصلی در حوزه تشخیص تقلب، نامتعادل بودن مجموعه دادگان است. Joo-Hyuk Oh و همکارانش در سال ۲۰۱۹ روشی را ارائه کردند که در آن با کمک روش یادگیری عمیق GAN به ایجاد نمونه‌های بیشتر پرداختند تا مشکل نامتعادل بودن داده‌ها را حل کنند و ناهنجاری‌ها را در دادگان تشخیص دهند. روش SMOTE که پیش‌تر برای حل مشکل نامتعادل بودن داده‌ها در نظر گرفته می‌شد در این پژوهش استفاده شد که برای نمونه‌گیری از داده‌هاست. اما مشکل اصلی این روش این بود که احتمال اینکه داده‌ها تولید شده به صورت بایاس شده باشند و به کارایی مدل طبقه‌بند ایجاد شده تأثیر بگذارند بسیار زیاد بود. برای غلبه برای این مشکل رویکرد cGAN ارائه شده است که احتمال اکثریت یا اقلیت داده‌ها را در نظر نمی‌گیرد و داده‌های بایاس تولید نمی‌کنند. روش ارائه‌شده توسط این پژوهش که OD-GAN نام دارد و با استفاده از معماری شبکه عمیق GAN که دارای دو بخش است به افزایش کارایی طبقه‌بند بین داده‌های نامتعادل با وجود داده‌های پرت در کلاس اکثریت (داده‌های هنجار) می‌پردازد. با تولید داده مصنوعی که توزیع مشابه داده‌های کلاس اقلیت (داده‌های پرت) به حل مشکل نامتعادل بودن داده‌ها می‌پردازد. سپس با استفاده از معماری شبکه GAN به

تشخیص داده پرت می‌پردازد. یکی از مشکلات اساسی این رویکرد پیچیدگی زمانی آموزش شبکه برای ایجاد داده مصنوعی جدید است[۴۹].

در مقاله‌ای که توسط Yu-Jun Zheng و همکارانش در سال ۲۰۱۸ انجام شده است، با استفاده از شبکه یادگیری عمیق GAN به تشخیص تقلب در داده‌های یک بانک چینی پرداخته‌اند. ایده اصلی مقاله توسعه یک روشی جهت شناسایی انتقال‌های مشکوک و متقلبانه بین فرستنده و گیرنده (در یک ارتباط از راه دور) در بانک پذیرنده تراکنش است. برای هر تراکنشی یک مقدار احتمالی را در نظر می‌گیرد و با تعیین یک حد آستانه به جلوگیری از انتقال تراکنش و شناسایی الگوی متفاوت و مشکوک بودن تراکنش می‌پردازد. در این پژوهش از روش یادگیری عمیق Deep Autoencoder Denoising برای مدیریت و تحلیل نویزهای ورودی استفاده شده و سپس از دو لایه طبقه‌بند برای افزایش کارایی و تأثیر بیشتر یادگیری استفاده شده است. با استفاده از شبکه GAN به استخراج ویژگی‌های پنهان ورودی‌های اولیه می‌پردازد[۵۰].

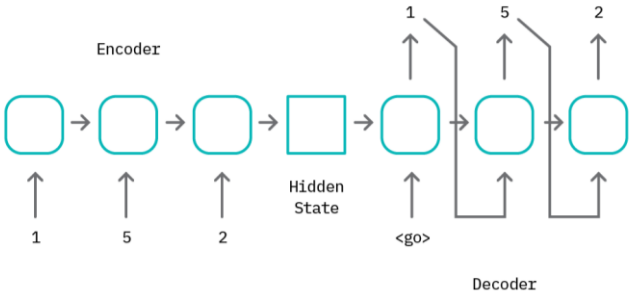
۲-۴-۴ مدل‌های رشته‌به‌رشته[1]

این روش‌ها کلاسی از شبکه‌های عصبی مصنوعی هستند که اساسا برای یادگیری نگاشت داده‌ها به صورت رشته‌ای از داده‌ها طراحی شده‌اند. داده‌هایی که به صورت رشته هستند به دلیل وابستگی توکن‌ها و ارتباط آن‌ها به هم چالش برانگیزند. برای مثال در حوزه پردازش زبان‌های طبیعی و بحث ترجمه ماشینی که یک رشته از کلمات در یک زبان به‌خصوص به ماشین داده می‌شود و نیاز دارد که نگاشت این رشته از کلمات در زبان متفاوت دیگر را به دست آورد. برای فائق آمدن به چنین مسائلی باید از مدلی استفاده کرد که محل هر توکن یا کلمه را در رشته در جملات بزرگتر به درستی آموزش ببینید و ارتباط بین رشته کلمات را حفظ کند.

همانند روش‌های دیگر یادگیری عمیق، مدل‌های رشته‌به‌رشته نیز از رمزگذار و رمزگشا استفاده میکند

[1] Sequence-to-Sequence Models

و یک بازنمایی از توکن ورودی را تولید می‌کند که وظیفه‌ی رمزگشا تولید رشته توکن‌ها خروجی توسط این بازنمایی ایجاد شده است. روش حافظه کوتاه مدت بلند[1] از روش‌های یادگیری عمیق رشته‌به‌رشته مهم است. برای شناسایی داده ناهنجار با استفاده از این رویکرد با کمک روش نیمه نظارتی به آموزش مدل رشته‌به‌رشته بر روی داده‌های نرمال می‌پردازیم. سپس با مقایسه (میانگین خطای بازسازی) بین رشته تولید شده نتیجه و محاسبه اختلاف آن‌ها می‌توان داده‌ها ناهنجار را شناسایی کرد.



شکل ۰-۶: معماری مدل‌های رشته به رشته

محققین با ارائه رویکردی جدید در داده‌ها رشته‌ای با استفاده از روش LSTM شبکه‌های مصنوعی عمیق بازگشت‌پذیر و مکانیزم توجه برای شناسایی تقلب پرداختند. ایده اصلی این مقاله توجه به ماهیت رشته‌ای بودن داده‌های تراکنش مالی در رشته‌های ورودی بوده است. در مقاله برای یادگیری بهتر داده‌های و عملکرد دقیق‌تر طبقه‌بند در مرحله اول از انتخاب ویژگی و الگوریتم‌های کاهش بُعد استفاده کردند. سپس برای حل مشکل عدم‌تعادل داده‌ها از روش نمونه‌برداری کمی ساختگی بیش از حد اقلیت مصنوعی (SMOTE)[2] استفاده کرده و سپس با مکانیزم توجه در شبکه‌های LSTM به یادگیری الگوی رفتار خرید مشتریان در رشته‌های تراکنش آن‌ها پرداختند. مجموعه دادگان این تحقیق شامل دو بخش بوده؛ بخش اول اطلاعات تراکنش‌های دارندگان کارت اعتباری اروپایی در سال ۲۰۱۳ به مدت دو روز در ماه سپتامبر و مجموعه دادگان بعدی شامل ۵۹۴۶۴۳ تراکنش ساختگی در طی ۱۸۰ روز بوده است.

[1] Long Short-Term Memory(LSTM)
[2] Synthetic Minority Over-sampling TEchnique

نتایج حاصل از اجرای الگوریتم حاکی از عملکرد و دقت نسبتاً بالای این رویکرد جدید بوده است[۵۱].

۲-۵ روش‌های مبتنی بر تحلیل گراف (گراف‌کاوی)

پیشرفت در ارتباطات و فنّاوری‌های دیجیتالی باعث به وجود آمدن دنیای به هم مرتبط از داده‌ها گشته است. انواع مختلف شبکه‌های ارتباطی مانند شبکه‌های بانکی، شبکه‌های تجارتی صنعتی، شبکه‌های اجتماعی در اینترنت و تارنماهای تجارت آنلاین از نمونه‌های ارتباط داده‌ها در قالب یک شبکه یا گراف در ریاضیات است.

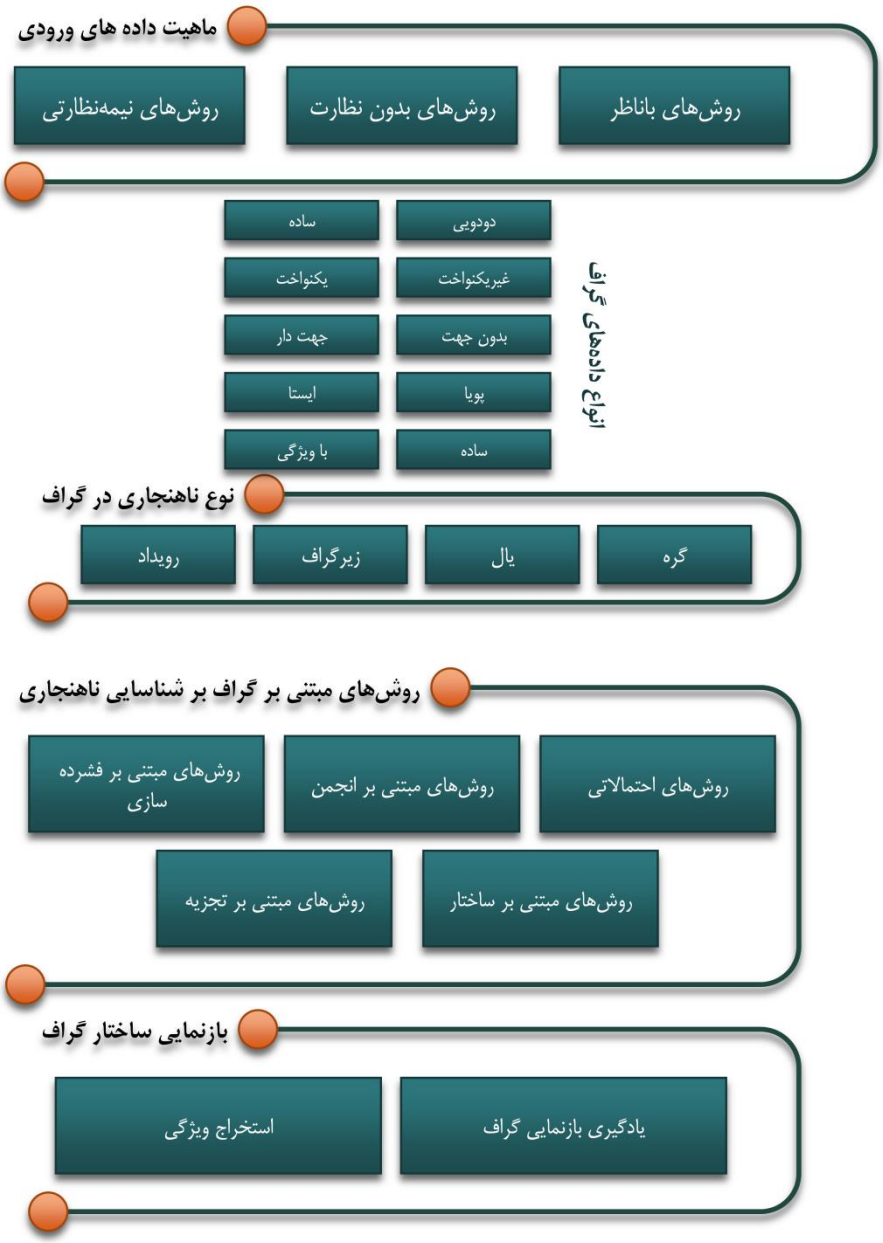
به دلیل ماهیت رابطه‌ای بودن داده‌ها، استفاده از تحلیل داده‌های شبکه‌های اجتماعی در راستای شناسایی داده‌ها ناهنجار محبوبیت زیادی بین محققین پیدا کرده است. شناسایی روابط و الگوهای ارتباطی بین داده‌های شبکه برای با کمک تحلیل شبکه‌های گرافی در سال‌های اخیر بسیار مورد توجه قرار گرفته است. اهمیت و دلیل استفاده از روش‌های استفاده از روش‌های تحلیل گراف بدین شرح است:

- ارتباط و وابستگی داده‌های دنیای واقعی نسبت به هم: بسیاری از داده‌ها در دنیای واقعی یک وابستگی درونی باهم دارند که به کمک آن می‌توان در شبکه‌های به‌هم‌پیوسته بزرگ، داده‌های ناهنجار را پیدا کرد. به عنوان مثال شبکه‌های خرده‌فروشی، شبکه‌های اجتماعی اینترنتی، شبکه تعاملی پروتئین-پروتئین و ... دارای به هم پیوستگی زیادی در بین گره‌های شبکه می‌باشند.
- ماهیت رابطه‌ای بودن داده ناهنجار در مجموعه دادگان دنیای واقعی: در واقع ماهیت داده ناهنجار نشان‌دهنده‌ی رابطه‌ای بودن آن است. به عنوان مثال در حوزه تقلب در تراکنش‌های مالی، متقلبین در بسیاری از موارد باهم در ارتباط هستند و شبکه‌ای را تشکیل می‌دهند. که این امر به کمک گراف و گراف‌کاوی به راحتی می‌توان نمایش داد.
- نمایش درست و قابل‌درک از داده‌های دنیای واقعی در قالب گره و یال‌های گراف: معیارها و اطلاعات مفیدی را از گراف‌ها می‌توان استخراج کرد که در تحلیل شبکه‌های گرافی بسیار حائز

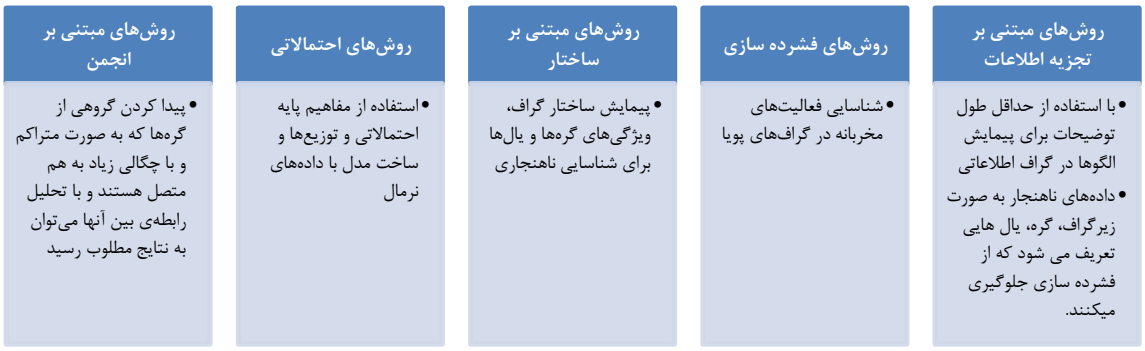
اهمیت است.

- استفاده از گراف و روش‌های تحلیل شبکه‌های گرافی باعث کمتر شدن دخالت انسان برای تحلیل داده و استفاده بهتر از تکنیک‌ها یادگیری ماشینی شده است.
- به دلیل برخط بودن داده‌های در دنیای امروز و رشد روزافزون داده‌ها می‌توان با استفاده از گراف‌های پویا تا حد زیادی این مشکل را برطرف کرد و از ویژگی‌های گراف پویا برای تحلیل و بررسی رفتار مشکوک داده‌ها در گراف‌ها استفاده کرد.

در شکل ۲-۷ یک چارچوب کلی از مطالعات در حوزه گراف را و پیش‌فرض‌های مورد نیاز را مشاهده می‌کنیم.



شکل ۰-۷: چارچوب کلی تشخیص ناهنجاری با گراف



شکل ۰-۸: روش‌های تشخیص ناهنجاری در گراف

اخیراً استفاده از روش‌های یادگیری بازنمایی گراف[1] (تعبیه گراف[2]) محبوبیت زیادی پیدا کرده است. این تکنیک به صورت خودکار ساختار داده‌های شبکه را رمزگذاری می‌کند. در واقع ایده اصلی این رویکرد یادگیری نگاشتی است که گره‌ها، یال‌ها و یا کل ساختار گراف را به فضای برداری با ابعاد کمتر تبدیل می‌کند تا ویژگی‌های ساختاری مهم پنهان را استخراج کند. این روش با مهندسی ویژگی متفاوت است. در شبکه‌های بزرگ و پویا که در طول زمان تغییر می‌کنند، روش مهندسی ویژگی بسیار هزینه‌بر، گران است و علاوه بر این نیز در مواردی که رفتار و الگوی مخرب داده‌ها در حال تغییر است مقیاس‌پذیری شناسایی مدل نیاز به به‌روزرسانی دارند که روش‌های مهندسی ویژگی چنین توانایی ندارند [۵۲].

تکنیک‌های یادگیری بازنمایی گراف برای استخراج اطلاعات ساختاری از شبکه‌ها بدون نیاز به دانش متخصصین در مورد داده‌ها است. این رویکرد در تشخیص تقلب در تراکنش‌های مالی برای شناسایی الگوهای مخرب در گراف بسیار کاربرد دارد. امروزه با کمک پیشرفت یادگیری عمیق، روش‌ها و تکنیک‌ها زیادی برای یادگیری بازنمایی گراف وجود دارد که با حجم زیاد داده‌ها می‌تواند به درستی عمل کند. در ادامه به بررسی پژوهش‌های انجام شده در این زمینه می‌پردازیم.

ولاسکو و هکارانش و در سال ۲۰۲۰ با استفاده از ترکیب روش‌های تئوری شبکه‌های اجتماعی و گراف، روش‌های بدون نظارت مانند خوشه‌بندی و تحلیل رگرسیون به شناسایی الگوی مشکوک و افرادی که در

[1] Graph Representation Learning
[2] Graph Embedding

مناقصه‌ها تبانی کرده اند پرداختند. در این تحقیق ابتدا و پس از جمع‌آوری داده‌ها از منابع مختلف و یکپارچه کردن آن‌ها با استفاده از گراف، داده‌ها در قالب گراف نمایش دادند، سپس با تحلیل بر روی گراف و استخراج اطلاعات مفید به خوشه‌بندی گره‌ها پرداختند. در ادامه با تعریف الگوهای تبانی مختلف (در سطح شرکت‌ها و در سطح افراد) و استفاده از الگوریتم خوشه‌بندی فضایی مبتنی بر چگالی به خوشه‌بندی داده‌ها پرداختند. داده‌های این تحقیق شامل معاملات و مناقصات کشور برزیل و اطلاعات شرکت‌ها و مناقصه‌گزاران و مناقصه‌گران بوده است که از سال ۲۰۱۸ جمع‌آوری شده است. نتایج حاصل از تحقیق نشان‌دهنده شناسایی نسبتاً مناسبی از شرکت‌ها و افرادی است که مشکوک بوده و در حال تبانی با دیگر شرکای خود بوده‌اند[۵۳].

ژو و همکارانش در سال ۲۰۲۰ تحقیقی را در مورد شناسایی تبانی بین شرکت‌ها در مناقصه‌های سال‌های ۲۰۱۱ تا ۲۰۱۸ در حوزه ساخت و ساز در کشور چین را انجام دادند. افراد مشکوک در مناقصات در سه سطح بررسی شده‌اند و رفتار مشکوک و تبانی شرکت‌ها را شناسایی کردند. ایده اصلی در این مقاله شناسایی انجمن‌های پرارزش در شبکه گراف و سپس بررسی ویژگی‌های هر انجمن در سطوح مختلف بوده است. ابتدا با استفاده از الگوریتم‌های شناسایی و جداسازی انجمن به تفکیک و خوشه‌بندی داده‌ها شبکه گراف پرداختند و در مرحله بعد با استفاده تحلیل معیارهای موجود در شبکه گرافی (معیار نزدیکی، معیار بینابینی و ضریب خوشه‌بندی) به شناسایی احتمال تبانی بین شرکت‌ها در مناقصات پرداختند. مجموعه داده‌ها تحقیق شامل ۳۱۵۰ گره و ۴۱۳۰۲ یال بوده است و نتایج حاصل تحقیق را در سه دسته‌ی مشکوک در سطح بالا، متوسط و کم دسته‌بندی کردند و حاکی از عملکرد مناسب رویکردهای تحلیل گرافی در داده‌ها بوده است[۵۴].

استفاده از تکنیک‌ها گراف‌کاوی و پیدا کردن بسترهای فساد تحقیقی است که وَن ارون و همکارانش در کشور برزیل بر روی معاملات دولتی انجام دادند. محققین در این مقاله با استفاده از روش‌های گراف‌کاوی و ابزار Neo4j به تحلیل گراف حاصل از تدارکات و خریدوفروش دولت برزیل پرداختند.

بسیاری از تقلب‌ها و فساد در خریدوفرش و تدارکات دولتی به صورت ارتباط گروهی از شرکت‌ها با هم صورت می‌گیرد و شرکت‌های که قصد تبانی و فساد را دارند با دستکاری فرآیند مناقصات و یا مزایده‌ها به نیت خود می‌رسند. در این تحقیق با بررسی و شناسایی رفتار و عملکرد شرکت و مطالعه ویژگی‌ها مشترک در بین شرکت‌ها متقلب به شناسایی تبانی پرداختند. ایده اصلی این مقاله استفاده از ابزار Neo4j و انجام جستجو بر روی گراف دانش بوده است[۵۵].

در سال ۲۰۱۶ ساروات و همکارانش روشی را برای شناسایی گروه‌های تبانی‌کننده در بازار سهام با استفاده از روش خوشه‌بندی طیفی[1] گرافی ارائه دادند. ایده اصلی این مقاله استفاده از روش‌های خوشه‌بندی طیفی در گراف و تعریف معیار نزدیکی و بینابینی بین دو معامله‌کننده و محاسبه حجم، تعداد و قیمت سهام معامله شده بین آن‌ها و استفاده از معیار فرضی‌ای برای امتیازدهی به گروه‌های مشکوک، به شناسایی گروه‌های با الگوی متفاوت در بازار سهام پرداختند. الگوریتم ارائه شده بر روی داده‌های واقعی SEBI اجرا شده و نتایج حاصل از نشان از عملکرد مناسب این الگوریتم است[۵۶].

با توجه به افزایش محبوبیت وب‌سایت‌های مزایده آنلاین، نگرانی‌ها در مورد تقلب و فعالیت مخربانه در این سایت‌ها افزایش یافته است. از آنجایی که بسیاری از روش‌های تقلب در حال تغییر هستند و روش‌های زیادی برای شناسایی این گونه تقلب‌ها وجود ندارد، دادفرنیا و همکارانش رویکردی برای شناسایی تبانی‌ها در شبکه‌های بزرگ مزایده آنلاین ارائه دادند. ایده اصلی این پژوهش استفاده از روش نیمه‌نظارتی ICAFD بوده است. ICAFD چهار مرحله دارد؛ استخراج ویژگی، محاسبه امتیاز ناهنجاری و مشکوک بودن داده، تشخیص تقلب و به روز رسانی افزایشی با استفاده از مفهوم انتشار باور. عملکرد مناسب این روش از لحاظ زمانی اجرا بر روی مجموعه دادگان از ویژگی‌های مثبت رویکرد ارائه شده است. نتایج حاصل از اجرای این روش حاکی از عملکرد بالای ICAFD است[۵۷].

در تحقیقی [۵۸] نویسندگان از روشی جدید برای یادگیری بازنمایی گراف‌های دودویی تراکنش‌های کارت اعتباری پرداختند. بر روی دو مجموعه داده با تعداد بالای گره و یال، این روش آموزش دیده و

[1] Spectral Clustering

نتایج حاصل از آن نشان‌دهنده‌ی مؤثر بودن روش تعبیه گراف در کاربرد پیش‌بینی لینک است. بردار ویژگی به دست آمده از روش تعبیه گراف، به عنوان ورودی الگوریتم‌های یادگیری ماشینی برای کاربردهای بیشتر و انجام عمل‌های دیگر داده می‌شود. از آنجا که تحقیقات بیشتر در مورد گراف‌های مشابه بود، نوآوری اصلی این مقاله کار بر روی گراف‌های غیرمشابه دودویی بوده است.

محققین در [۵۹] با استفاده از خوشه‌بندی مبتنی بر تعبیه عمیق داده‌ها به یادگیری بازنمایی ویژگی‌ها و ارزیابی خوشه‌بندی تواما پرداختند. این روش که به طور مختصر DEC نامیده می‌شود ابتدا با یادگیری نگاشت داده‌ها به فضای ویژگی با ابعاد کوچک‌تر و سپس بهینه‌سازی تابع هدف خوشه‌بندی می‌پردازد. مجموعه داده این تحقیق بر روی عکس MNIST و داده‌های متنی اخبار رویترز بوده و نتایج حاصل از عملکرد مناسب این رویکرد بوده است و ضمناً این روش به انتخاب فرا پارامترها حساسیت کمتری نسبت به دیگر روش‌ها نشان داده است.

برخلاف بسیاری از روش‌هایی که به شناسایی گره‌های مشکوک در شبکه برای شناسایی تقلب در تراکنش‌های مالی می‌پردازند، یولونگ پی و همکاران شناسایی زیرگراف‌های مشکوک شبکه‌های تراکنش‌های مالی را مورد تحقیق قرار دادند [۶۰]. چارچوب SADE ارائه شده در این تحقیق از دو مرحله تشکیل شده؛ مرحله اول تعبیه گراف با نقش‌های هدایت شده و مرحله دوم شناسایی زیرگراف‌های مشکوک. نویسندگان در این تحقیق ادعا کرده اند که رویکرد ارائه شده هم در شناسایی زیرگراف‌های محلی و هم شناسایی زیرگراف‌ها در کل ساختار گراف کاربرد دارد. در واقع با ترکیب اطلاعات ساختاری از نقش‌های گره‌ها و ماتریس همجواری بین زیرگراف‌های مختلف با استفاده از تکنیک تعبیه گراف گام تصادفی به شناسایی زیرگراف‌های محلی و کل گراف پرداختند. مجموعه داده ساختگی و داده واقعی در این تحقیق استفاده شده است.

در تحقیقی [۶۱] که در سال ۲۰۱۹ انجام شد محققین با استفاده از تکنیک‌های شبکه‌های عصبی مصنوعی عمیق به یادگیری بازنمایی گراف در تراکنش‌های مالی شبکه‌های پرداخت دنیای پردختند. در

این تحقیق از الگوریتم شبکه کانولوشنی گراف برای یادگیری تعبیه کردن گره‌ها و یال‌های استفاده شده است. همانند دیگر تحقیق‌ها، بعد از ایجاد بردار ویژگی با ابعاد دیگر، از الگوریتم‌های یادگیری ماشین استفاده می‌شود. نتایج حاصل از تحقیق که بر روی دو مجموعه داده اجرا شده است، حاکی از دقت بالای این روش در طبقه‌بندی گره‌ها با دقت ۹۷,۳۴ درصد بوده است و از دیگر روش‌ها مانند جنگل تصادفی و GraphSAGE عملکرد بهتری داشته است.

همانطور که گفته شد، یکی از چالش‌ها در زمینه شناسایی داده متقلبانه، عدم تعادل و توازن بین داده‌ها و مجموعه دادگان است. بدین معنی که در یک مجموعه داده تعداد داده‌های ناهنجار بسیار کمتر است و باعث کاهش عملکرد و دقت شناسایی سیستم تشخیص تقلب می‌شود. در همین راستا لیو و همکارانش در سال ۲۰۲۰ با استفاده مدل‌های مبتنی بر گراف به شناسایی افراد مشکوک در سامانه‌ها برخط پرداختند. با توجه به پیشرفت و اهمیت زیاد شبکه‌های عصبی گرافی، در این مقاله از این روش برای شناسایی افراد مشکوک در گراف‌های یکنواخت و غیریکنواخت استفاده شده است. ایده اصلی روش‌های شبکه عصبی گرافی جمع‌آوری اطلاعات همسایه‌های هر گره برای یادگیری تعبیه هر گره است. فرض اصلی در شبکه عصبی گرافی این است که همسایه‌های هر گره اطلاعات مشابه، ویژگی‌ها و روابط یکسان دارند. الگوریتم ارائه شده در این مقاله برای حل مشکل عدم تعادل و توازن بین داده‌های هر گره و همسایه‌های آن است که با استفاده از چارچوب شبکه عصبی گرافی و الگوریتم GraphConsis پرداختند. نمونه برداری از همسایه‌ها و امتیاز به پایداری هرکدام از آن گره‌ها از کارهای اصلی این تحقیق بوده است. مجموعه دادگان این تحقیق نظرات اسپم YelpChi است و نتایج حاصل از تحقیق حاکی از دقت بالا این الگوریتم ارائه شده است و نسبت به دیگر روش‌ها عملکرد بهتری داشته است[۶۲].

تشخیص تقلب در داده‌های برخط که جریان داده همیشه برقرار است با استفاده از رویکرد شبکه عصبی گرافی غیریکنواخت موضوع تحقیقی است که وانگ و همکارانش در سال ۲۰۲۱ به آن پرداختند. با توجه به افزایش تعداد فروشگاه‌های آنلاین و پرداخت‌های آنلاین و داده‌های برخطی که در هر لحظه در

حال انتقال هستند، وجود روشی برای شناسایی رفتار متقلبانه مشتریان در این پلتفرم‌ها به دلیل ناهمگون بودن شبکه گرافی مسئله‌ی مهمی است. در این مقاله با ارائه رویکردی نوآورانه که به اختصار LIFE نامیدند به یادگیری رفتار و اطلاعات در مورد تراکنش‌ها پرداختند. در این رویکرد از الگوریتم انتشار برچسب برای حل مشکل تعداد کم نمونه‌های متقلبانه آموزش دیده استفاده کردند. مجموعه دادگان این تحقیق از پلتفرم Taobao که در دنیای واقعی وجود دارد استفاده کردند. نتایج حاصل از تحقیق حاکی از دقت ۹۶ درصدی این روش در مقایسه با دیگر روش‌ها است [۶۳].

۲-۶ جمع‌بندی

در این فصل به بررسی تحقیق‌های موجود در زمینه تشخیص تقلب در تراکنش‌های مالی پرداختیم. ابتدا رویکردهای یادگیری ماشین و داده‌کاوی را مورد بررسی قرار دادیم. سپس تکنیک‌های یادگیری عمیق مورد بحث و بررسی قرار گرفت. در انتها نیز به دلیل اهمیت و مشابهت این پژوهش با روش‌های تحلیل شبکه‌های اجتماعی و گراف‌کاوی رویکردهایی که از تحلیل گراف و ترکیب آن با یادگیری عمیق بوده را مورد مطالعه قرار دادیم. در فصل آینده به ارائه راهکار پیشنهادی و نتایج و ارزیابی‌های آن می‌پردازیم.

فصل ۳ رویکرد پیشنهادی

۳-۱ مقدمه

همانطور که در فصول پیش گفته شد یکی از مهم‌ترین کاربردهای تشخیص ناهنجاری، در حوزه‌ی مالی است. تشخیص تقلب در مناقصات و معاملات و خریدوفروش‌ها از مسائلی است که کشف و شناسایی رفتار مشکوک افراد در معاملات باعث جلوگیری از ضررهای اقتصادی می‌شود. در ادامه و در فصل دو پژوهش‌هایی را که در این زمینه مطرح شد را نسبتاً تا حد زیادی بررسی کردیم و نشان دادیم که استفاده از روش‌های یادگیری عمیق و تحلیل شبکه‌های گرافی معاملات در این حوزه به دلیل ماهیت رابطه‌ای بودن داده‌ها در حال افزایش است و محققین زیادی از ترکیب این مفاهیم برای شناسایی و کشف رفتار و الگوها مخرب در داده‌ها استفاده می‌کنند.

در این فصل برای روشن‌تر شدن موضوع مورد بحث ابتدا به توضیح در مورد مجموعه دادگان می‌پردازیم و در ادامه معماری پیشنهادی و رویکرد استفاده شده در این پژوهش را مورد بررسی قرار می‌دهیم. سپس بخش‌های مختلف راه‌حل ارائه شده را شرح خواهیم داد و مفاهیم و پیش‌نیازها را بررسی می‌کنیم. راه‌حل ارائه شده در این پژوهش (که به صورت کاربردی است) ترکیبی از گراف‌کاوی، تحلیل شبکه‌های اجتماعی، روش‌های یادگیری عمیق و تکنیک داده‌کاوی است. بیشتر تحقیقات انجام شده در این زمینه و با این رویکرد در سال‌های ۲۰۱۹ به بعد بوده و ترکیب این روش‌ها که در بعضی مقالات به آن اشاره شده حاکی از عملکرد مناسب و کاربردی بودن آن است.

۳-۲ مجموعه دادگان پژوهش

داده‌های این پژوهش که داده‌های واقعی هستند شامل اطلاعات مربوط به مناقصه‌گزاران (دستگاه‌های اجرایی) و مناقصه‌گران (تأمین‌کنندگان) در سطح کشور ایران است که در سامانه تدارکات دولت (ستاد) ثبت شده و در تارنمای سامانه[۱] در بخش شفاف‌سازی معاملات قرار داده شده، گرفته شده است. این سامانه توسط مرکز توسعه تجارت الکترونیکی وزارت صنعت، معدن و تجارت از سال ۱۳۸۸ به صورت یک

[۱]. به آدرس: https://setadiran.ir/

طرح ملی آغاز به کار کرده و با مصوبه هیئت دولت در سال ۱۳۹۰ تمام دستگاه‌های اجرایی ملزم به استفاده از این سامانه هستند. این سامانه به منظور انجام معاملات و عقد قرارداد (مناقصه، مزایده و خریدوفروش) و شفافیت‌های مالی دستگاه‌های اجرایی در بستر وب ایجاد شده است و تمام صاحبان مشاغل و کسب‌وکارهای گوناگون می‌توانند در معاملات دولتی برحسب نیاز و درخواست‌هایشان مشارکت نمایند.

در تحقیق پیش رو تعداد ۳۲۱۵۲ معامله مناقصه برگزار شده طی سال‌های ۱۳۸۹ الی تیر ۱۳۹۸ مورد تحلیل قرار گرفته است. مناقصه‌گزار دستگاه اجرایی‌ای است که تشریفات خرید را از طریق مناقصه برگزار می‌کند و مناقصه‌گر شخصی حقیقی یا حقوقی‌ای که اسناد مناقصه را دریافت می‌کند و در مناقصه برای فروش شرکت می‌کند. معاملات برگزار شده بین تأمین‌کننده و مناقصه‌گزار به صورت یال بدون جهت در نظر گرفته شده است و با بررسی‌های انجام‌شده دستگاه‌های مناقصه‌گزار که فقط یک معامله داشتند از ساختار گراف حذف گردید. در جدول ۳-۱ اطلاعات داده‌های مورد استفاده آورده شده است.

جدول ۰-۱: مجموعه دادگان مورد استفاده در تحقیق

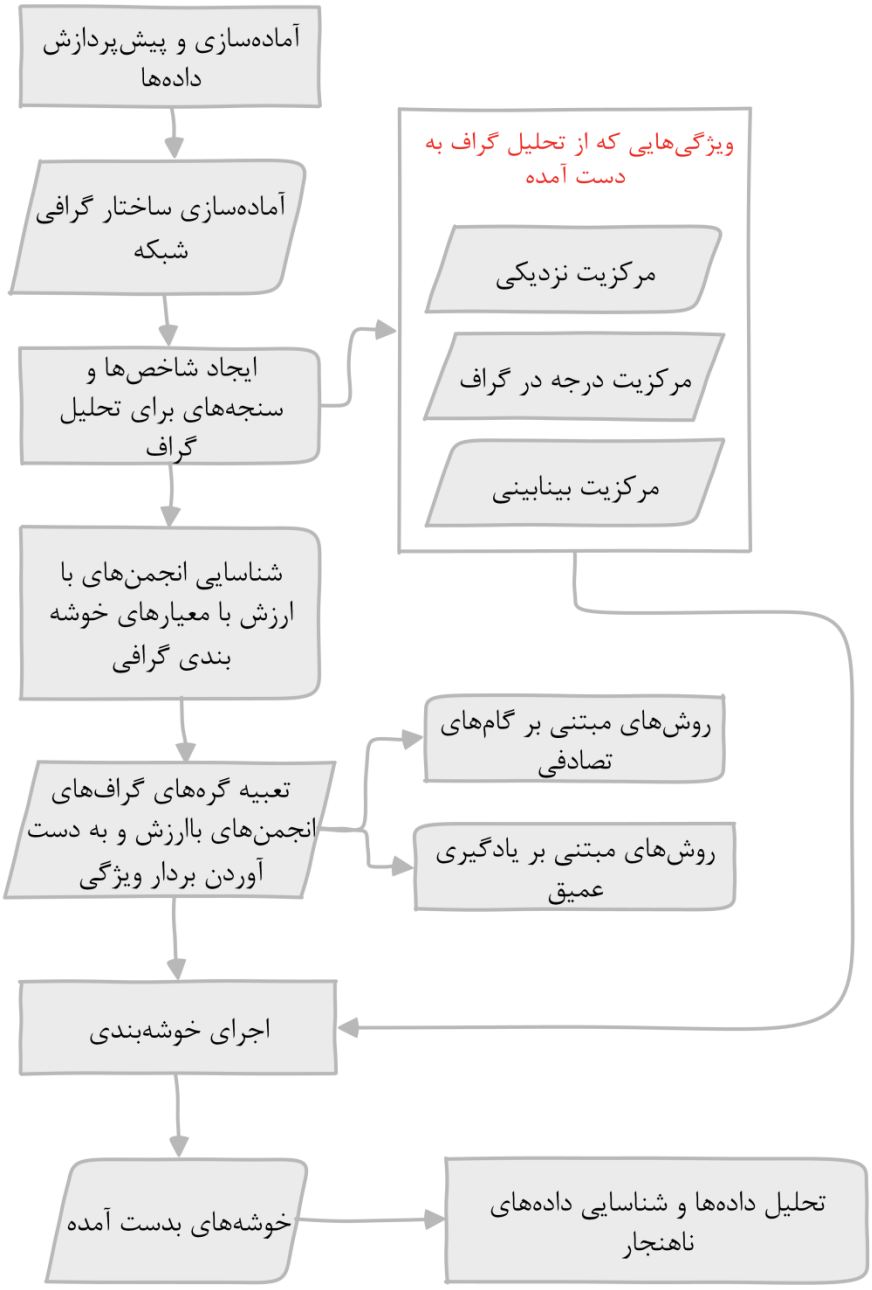
شرح	فراوانی(درصد)	نوع گره	تعداد
اطلاعات مناقصه گزاران	۱۲,۲	مناقصه‌گزار	۱۷۹۹
اطلاعات مناقصه‌گران	۸۷,۸	مناقصه‌گر	۱۲۹۷۰
	۱۰۰		۱۴۷۶۹

برای مناقصه‌گزاران و مناقصه‌گران ویژگی‌هایی وجود دارد که در ادامه به هریک اشاره خواهیم کرد.

- ویژگی‌های اطلاعات مناقصه‌گزاران: شماره شناسایی، نام دستگاه اجرایی، نام استان، کد استان، مبلغ معاملات، تعداد معاملات، شماره دستگاه اجرایی مرکزی.
- ویژگی‌های اطلاعات مناقصه‌گران: شماره شناسایی، نام، شماره و کد استان، نام استان، تعداد معاملات، مبلغ معاملات.

۳-۳ ساختار معماری و رویکرد پیشنهادی

در این بخش جزئیات هر مرحله از معماری پیشنهادی شکل ۳-۱ را توضیح می‌دهیم و به شرح بخش‌های مختلف و الگوریتم‌های ارائه شده می‌پردازیم.



شکل ۰-۱: شمای کلی معماری پیشنهادی و رویکرد ارائه شده

۳-۳-۱ پیش‌پردازش و آماده‌سازی ساختار گرافی اطلاعات

در مرحله اول بعد از انجام پیش‌پردازش بر روی داده‌ها، با استفاده از تکنیک‌های تحلیل شبکه‌های گرافی و گراف‌کاوی به آماده‌سازی و ایجاد ساختار گرافی اطلاعات پرداختیم. اطلاعات مورد نیاز معاملات انجام شده بین مناقصه‌گران و مناقصه‌گزاران با انجام جستجو و کوئری‌های مختلف در پایگاه داده انجام شده و اطلاعات خام آن‌ها را دریافت کرده‌ایم. برای انجام پیش‌پردازش بر روی داده‌ها ابتدا داده‌ها و ویژگی‌های که از اهمیت بالایی برخوردار بودند را شناسایی کرده و سپس بعد ازمدل‌سازی گرافی داده‌ها، برای هر گره ویژگی‌هایی را تعریف اختصاص دادیم.

همانطور که در شکل ۳-۲ مشاهده می‌کنیم در ستون اول شناسه‌های دستگاه‌های اجرایی هستند و در ستون دوم دستگاه‌های تأمین‌کننده قرار دارند. ویژگی‌هایی همچون تعداد معاملات، ارزش ریالی معاملات، شناسه‌ی استان تأمین‌کننده، نام استان تأمین‌کننده و شناسه‌ی دستگاه اجرایی و نام استان دستگاه اجرایی. لازم به ذکر است که نام دستگاه اجرایی و تأمین‌کننده نیز در جداول مربوط به هر کدام از آن‌ها نیز موجود است. یعنی در جداول اطلاعات شامل دستگاه‌های اجرایی و جداول اطلاعات شامل تأمین‌کنندگان.

دستگاه اجرایی	تامین‌کننده	تعداد معاملات	ارزش ریالی معاملات	شناسه و نام استان تامین کننده		شناسه و نام استان دستگاه اجرایی	
6884	666128	2	3067501720	420	کرمانشاه	420	کرمانشاه
6884	674934	4	38024225917	420	کرمانشاه	420	کرمانشاه
6901	681873	1	6416843427	420	کرمانشاه	420	کرمانشاه
4869	731938	1	6054486247	420	کرمانشاه	420	کرمانشاه
6915	720543	1	4700000000	420	کرمانشاه	420	کرمانشاه
6952	715888	1	8213504328	420	کرمانشاه	420	کرمانشاه
6915	653535	1	3250000000	401	آذربایجان غربی	420	کرمانشاه
6074	747436	1	8150000000	406	تهران	420	کرمانشاه
8678	725523	1	45594750872	420	کرمانشاه	420	کرمانشاه
6915	708869	1	8320728184	420	کرمانشاه	420	کرمانشاه
6086	683515	2	17655027364	420	کرمانشاه	420	کرمانشاه

شکل ۳-۲۰: بخشی از داده‌ها که به صورت گراف هستند با ویژگی‌های آن‌ها

برای مدل‌سازی گرافی، دستگاه اجرایی و تامین‌کننده هریک به عنوان گره‌های گراف در نظر گرفته می‌شوند و ارتباط بین گره‌های نیز تعداد معاملات و ارزش ریالی معاملات است که در ادامه و در فصل چهار در مورد گراف معاملات و ویژگی‌هایی که برای گره‌ها در نظر گرفته‌ایم بیشتر پرداخته‌ایم.

بعد از آشنایی و فهم در مورد داده‌ها، برای تحلیل گراف شبکه‌ی به دست آمده باید شاخص‌ها و معیارها را در هر گراف ساختاری به دست بیاوریم تا اطلاعات مفید در مورد گراف را دریابیم. در جدول ۳-۲ سنجه‌ها و معیارهای مختلف در شبکه‌های اجتماعی را مشاهده می‌کنیم.

جدول ۳-۲: سنجه‌ها و معیارهای تحلیل شبکه‌های اجتماعی

سنجه‌ها	توضیح	معیارها
ارتباطات	شباهت میان گره‌ها	یکریختی[1]، رابطه متقابل، چندگانگی[2]، بسته بودن شبکه و نزدیکی بین آن‌ها
توزیع‌ها	پراکندگی و ارتباطت میان شبکه‌ها و گره‌ها	پل، مرکزیت‌های نزدیکی و بینابینی، چگالی شبکه گرافی و چاله ساختاری
بخش‌بندی	برای یافتن بخش‌ها و خوشه‌های گوناگون در گراف	ضریب خوشه‌بندی و ماژولاریتی[3]

مرکزیت درجه[4]:

مهم‌ترین و ساده‌ترین معیار که مشخص‌کننده تعداد یال‌های متصل به هر گره است. با استفاده از این معیار می‌توان دریافت که یک گره دقیقاً با چند یال به دیگر قسمت‌های شبکه گرافی متصل است. از ین معیار برای شناسایی گره‌های مهم و با حجم اتصال بالا استفاده می‌شود تا در مورد آن اطلاعات ارزشمند را به دست بیاوریم. در گراف‌های جهت‌دار این معیار به دو صورت است که برای هر گره‌های تعداد یال‌های ورودی و تعداد یال‌های خروجی را محاسبه می‌کند.

مرکزیت نزدیکی[1]:

[1] Homophily
[2] Multiplexity
[3] Modularity
[4] Degree Centrality

معیاری برای هر گره که میزان نزدیکی به دیگر نودها را اندازه‌گیری می‌کند. بدین معنی که هر گره چه تعداد همسایه دارد و با همسایگان خود چه میزان نزدیک است. این معیار کوتاه‌ترین مسیر بین همه‌ی گره‌ها را محاسبه می‌کند و به هر گره امتیازی بر اساس مجموع کوتاه‌ترین مسیرها می‌دهد. از این معیار برای شناسایی محل مناسب گره تاثیرگزار روی دیگر گره‌ها استفاده می‌شود. در یک گراف هر چه میزان این معیار بالاتر باشد، نشان‌دهنده‌ی دسترسی به دیگر گره‌ها توسط این گره است.

برای محاسبه این معیار برای هر گره از فرمول زیر استفاده می‌شود. در فرمول ۳-۱ فرض می‌شود که d_{ij} فاصله بین کوتاه‌ترین مسیر بین دو گره j و i است، میانگین فاصله l_i به صورت فرمول ۳-۱ محاسبه می‌شود. از آنجایی که به دنبال گره‌های نزدیک‌تر هستیم، پیش معیار مرکزیت نزدیکی C_i به صورت معکوس میانگین طول مانند فرمول ۳-۲ محاسبه می‌شود.

$$l_i = \frac{1}{n}\sum_j d_{ij} \qquad (۳-۱)$$

$$C_i = \frac{1}{l_i} = \frac{n}{\sum_j d_{ij}} \qquad (۳-۲)$$

مرکزیت بینابینی[2]:

این معیار بیانگر جایگاه یک گره درون شبکه است و تعداد دفعاتی که هر گره بر روی کوتاه‌ترین مسیر بین گره‌ها قرار دارد را محاسبه می‌کند. این معیار نشان‌دهندی اهمیت یک گره در مسیر ارتباطی سایر گره‌ها است. این معیار نشان می‌دهد یک گره به صورت پل ارتباطی بین شبکه است یا خیر. محاسبه‌ی همه‌ی کوتاه‌ترین مسیرها و شمارش تعداد دفعاتی که هر گره در این کوتاه‌ترین مسیرها قرار دارد. از این معیار برای شناسایی گره‌های تأثیر گذار در گره برای انتشار اطلاعات استفاده می‌شود.

برای محاسبه معیار بینابینی برای هر گره برای هر نود مبدأ مانند s و نود مقصد مانند t و ورودی نود

[1] Closeness Centrality
[2] Betweenness Centrality

i که $s \neq t \neq i$ است و مقدار n_{st}^{i} یک است اگر نود i در بین کوتاه‌ترین مسیر بین نودهای s و t قرار بگیرد؛ در غیر این صورت صفر است. از آنجایی که بیشتر از یک کوتاه‌ترین مسیر بین دو نود s و t وجود دارد، بنابراین باید به کل تعداد کوتاه‌ترین مسیرها تقسیم کنیم و فرمول ۳-۴ برای محاسبه مرکزیت بینابینی درنهایت استفاده می‌شود.

$$x_i = \sum_{st} n_{st}^{i} \qquad (۳-۰)$$

$$x_i = \sum_{st} \frac{n_{st}^{i}}{g_{st}} \qquad (۴-۰)$$

۳-۳-۲ انجمن و شناسایی انجمن‌های باارزش در گراف

یکی از مهم‌ترین ویژگی‌های هر شبکه گراف، شناسایی وابستگی‌ها و انجمن‌های موجود در گراف است. انجمن شامل گروهی از گره‌ها است که شامل ویژگی‌های مشابه و الگوهای یکسان هستند و مانند خوشه‌بندی عمل می‌کند. برای شناسایی انجمن‌های موجود در یک گراف الگوریتم‌های مختلفی وجود دارد. یکی از الگوریتم‌های موجود ماژولاریتی است که به شناسایی الگوهای پنهان بین گره‌های می‌پردازد و به دنبال گروهی از گره‌هایی است که به صورت متراکم با هم در ارتباط هستند. همانند خوشه‌بندی یک شبکه گرافی با مقدار ماژولاریتی بالا ارتباط فشرده‌ای بین گره‌های درون یک انجمن و فاصله زیاد بین گره‌های در انجمن‌های دیگر است.

برای محاسبه ماژولاریتی داریم:

احتمال یال موجود در ماژول i — یال‌های موجود در ماژول i است e_{ii}

$$e_{ii} = |\{(u,v): u \in V_i, v \in V_i, (u,v) \in E\}|/|E|$$

احتمال یال تصادفی که در ماژول i قرار می‌افتد — یال‌هایی که حداقل یک سر آن‌ها در ماژول i است a_i

$$e_{ii} = |\{(u,v): u \in V_i, (u,v) \in E\}|/|E|$$

ماژولاریتی برابر است با:

$$Q = \sum_{i=1}^{k} (e_{ii} - a_i^2) \qquad (۵-۰)$$

۳-۳-۳ تعبیه گره‌های گراف

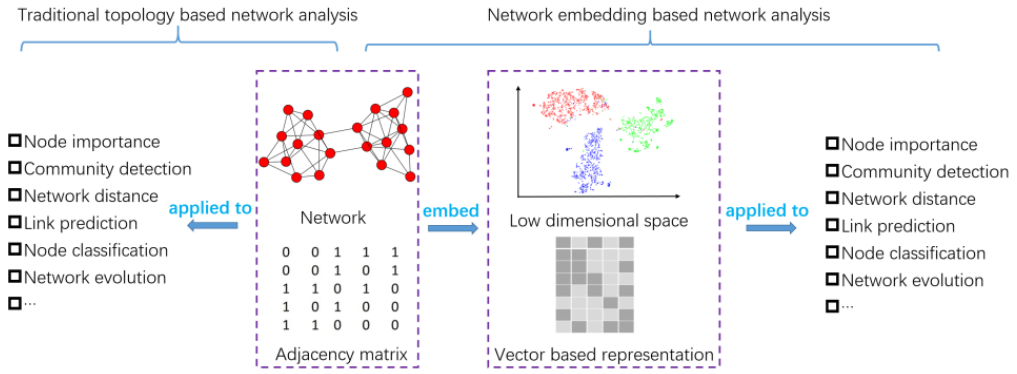
با توجه به اندازه گراف‌های شبکه‌های اجتماعی و تعداد بسیار زیاد گره‌ها و یال‌های بین آن‌ها، تحلیل شبکه‌های گرافی در بسیاری از کاربردهای در دنیای واقعی نقش مهمی را ایفا می‌کنند. تحلیل مفید و بهینه اطلاعات شبکه‌های گرافی بزرگ، وابستگی زیادی به نحوه‌ی نمایش و بازنمایی داده‌های گرافی شبکه‌ها دارد. در سال‌های اخیر حوزه یادگیری بازنمایی گراف باعث جذب محققین شده و تلاش‌های زیادی در این زمینه انجام شده است. هدف، یادگیری بازنمایی پنهان و با ابعاد کم داده‌هاست، درحالی که از ساختار گرافی شبکه، محتوای گره‌ها و دیگر اطلاعات جانبی استفاده می‌کند.

کار کردن با این‌گونه داده‌های گرافی برای اعمال الگوریتم‌های یادگیری ماشین و تکنیک‌های داده‌کاوی، پیچیده و چالش‌برانگیز است. برای اعمال این الگوریتم‌ها اولین کار پیدا کردن روشی مؤثر برای بازنمایی داده‌ها است. با نمایش داده‌ها به صورت دقیق می‌توان الگوها را کشف کرد، تحلیل و پیش‌بینی را بر روی داده‌ها انجام داد که در پیچیدگی زمانی و فضایی بهینه‌ای را داراست.

بسیاری از کارهای اولیه که در حوزه یادگیری بازنمایی داده‌های شبکه گرافی به سال‌های قبل از ۲۰۰۰ برمی‌گردد و محققین از تکنیک‌های کاهش بُعد استفاده می‌کردند. بدین صورت که یک مجموعه داده به عنوان ورودی داده شده (D – بُعد)، سپس با محاسبه میزان شباهت بین جفت نقطه‌هایی که گراف وابستگی را می‌سازند و سپس تعبیه کل گراف وابستگی که دارای ابعاد کمتری (d – بُعد) بود می‌پرداختند که در ان $d \ll D$ است. روش‌های ISOMAP، Locally Linear Embedding و Laplacian Eigenmap از روش‌هایی هستند که از چنین منطقی پیروی می‌کردند. مشکل اصلی این روش‌ها پیچیدگی زمانی $O(|V|^2)$ آن‌ها است که V تعداد گره‌های گراف است.

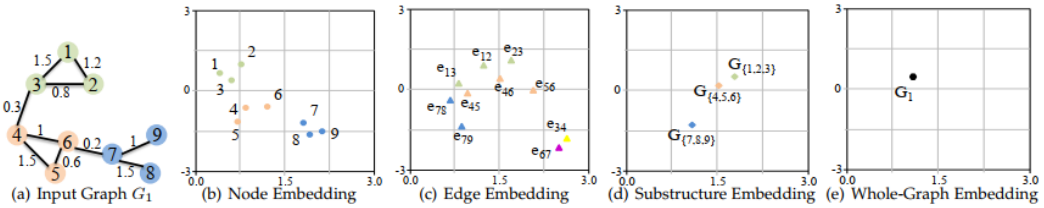
روش‌های پیشین برای بازنمایی داده‌های گرافی چالش‌های زیادی را در هنگام انجام فرآیند و تحلیل

آن‌ها ایجاد می‌کردند: حجم زیاد ماتریس همسایگی، پیچیدگی محاسباتی بالا، عدم توانایی برای اعمال الگوریتم‌های یادگیری ماشین.



شکل ۰-۳: مقایسه بین روش‌های تحلیل گرافی مبتنی بر ساختار شبکه و روش‌های مبتنی بر تعبیه گراف و شبکه[۶۴]

تعبیه گراف به معنی یادگیری بازنمایی برداری گره‌های گراف با ابعاد است. در فضای تعبیه‌شده گراف، ارتباط میان گره‌ها (که در گراف به صورت یال‌هایی متصل با دیگر گره‌ها و ساختارها نمایش داده می‌شود) به‌وسیله‌ی فاصله‌ی میان فضای برداری آن‌ها اندازه گرفته می‌شود و توپولوژی و ویژگی‌های ساختاری هر گره در بردارهای تعبیه‌شده کدگذاری می‌شوند. در واقع ایده‌ی اصلی روش‌های تعبیه گراف این است گره‌های نزدیک به هم در گراف در فضای برداری نیز نزدیک به هم هستند و ارتباط هندسی گره‌ها به هم در فضای برداری هم نزدیک به هم هستند. تعبیه گراف به چندین شکل انجام می‌شود که در شکل ۳-۵ مشاهده می‌کنیم. تعبیه گره، تعبیه یال، تعبیه زیر ساختار گراف و تعبیه کل گراف.



شکل ۰-۴: مثالی از تعبیه گراف در فضای دوبعدی و روش‌های مختلف تعبیه گراف (تعبیه گره، تعبیه یال، تعبیه زیر ساختار گراف و تعبیه کل گراف)[۶۵]

در این پژوهش نیز بعد از نمایش داده‌ها به شکل گراف، تحلیل گراف و کسب اطلاعات ارزشمند در مورد هر گره و یال‌ها و پیدا کردن انجمن‌های پرارزش، به یادگیری بازنمایی گره‌ها پرداختیم و خروجی

یادگیری بازنمایی حاصل از گره‌ها را به عنوان ویژگی اضافه به همراه دیگر ویژگی‌های داده‌ها به الگوریتم خوشه‌بندی برای تحلیل بیشتر بر روی داده‌ها و شناسایی خوشه‌ها و داده‌ها ناهنجار و با رفتار و الگوهای متفاوت می‌دهیم.

جدول ۰-۳: دسته‌بندی‌ای از الگوریتم‌های یادگیری بازنمایی شبکه‌های گرافی

متدولوژی	الگوریتم‌ها	ویژگی‌های مثبت	ویژگی‌های منفی
فاکتور سازی ماتریس	GraRep، HOPE، M-NMF	ساختار کلی گراف در نظر می‌گیرد	هزینه زمانی و محاسباتی بالا
گام‌های تصادفی	DeepWalk و Node2vec	نسبتاً بهینه در پیدا کردن مشابهت بین گره‌ها	فقط ساختارهای محلی را در نظر می‌گیرد
یادگیری عمیق	DNGR و SDNE	غیرخطی بودن روش‌ها	هزینه زمانی نسبتاً بالا

پیش از صحبت در مورد روش‌های مورد استفاده برای تعبیه گره‌های گراف باید به برخی از مفاهیم بپردازیم:

- همجواری مرتبه اول: مشابهت مستقیم بین گره‌ها با یکدیگر را در نظر می‌گیرد.

$$P_1(v_i, v_j) = \frac{1}{1 + \exp(-\vec{u}_i^T . \vec{u}_j)}$$

- همجواری مرتبه دوم: مشابهت بین گره‌های مشترک در همسایگی‌ها را در نظر می‌گیرد. همسایگی‌های دو گره را مقایسه می‌کند و مشابهت آن‌ها را می‌سنجد.

$$P_2(v_j|v_i) = \frac{\exp(\vec{u}_j^T . \vec{u}_i)}{\sum_{k=1}^{|V|} exp\,(\vec{u}_k^T . \vec{u}_i)}$$

در این پژوهش برای تعبیه گره‌های گراف از روش‌های DeepWalk [۶۶]، Node2vec [۶۷]و روش SDNE [۶۸]استفاده کردیم.

جدول ۰-۴: روش‌های تعبیه گراف استفاده شده در این پژوهش

روش‌های مبتنی بر گام‌های تصادفی	**روش DeepWalk**	استفاده از تکنیک Word2vec و گام‌های تصادفی از نودها	با پیچیدگی زمانی $O(\|V\|d)$ d ابعاد داده‌ها
	روش Node2vec	همانند روش DeepWalk به همراه استراتژی‌هایی برای نمونه‌برداری	
روش‌های مبتنی بر یادگیری عمیق و شبکه‌های خودرمزگذار	**روش SDNE**	استفاده از شبکه‌های خودرمزگذار و همجواری مرتبه اول و مرتبه دوم	با پیچیدگی زمانی $O(\|V\|\|E\|)$

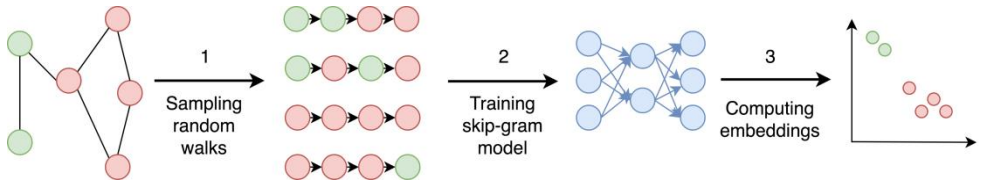
روش‌های که مبتنی بر گام‌های تصادفی برای تخمین ویژگی‌های مختلفی در گراف مانند مرکزیت و مشابهت در گراف استفاده می‌شوند. این روش‌ها بخشی از گراف (مخصوصاً در حالتی که گراف بسیار بزرگ باشد) را پیمایش می‌کنند.

۳-۳-۳-۱ روش $DeepWalk$

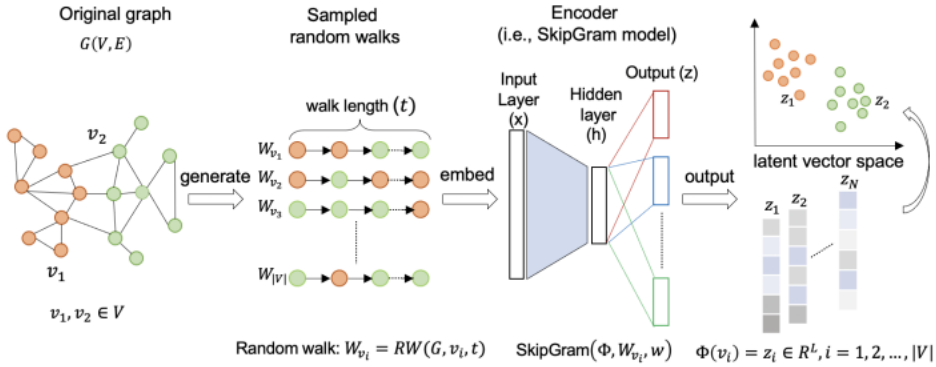
روش با استفاده از مجاورت مرتبه بالاتر بین گره‌ها و بیشینه‌سازی احتمال مشاهده‌ی آخرین k-گره و k-گره بعدی است که در پیمایش تصادفی مشاهده می‌شود. این مدل چندین گام تصادفی که طول هر یک از آن‌ها $2k+1$ است تولید می‌کند و برای بهینه‌سازی بر روی مجموع $\log - \text{likelihood}$ اجرا می‌شود. از یک رمزگشای مبتنی بر ضرب داخلی برای بازسازی یال‌ها از نودهای تعبیه‌شده استفاده می‌کند.

این روش به نوعی شبکه عصبی گرافیکی است که بر روی ساختار گراف اجرا می‌شود. از گام‌های تصادفی برای پیمایش گراف و جمع‌آوری اطلاعات ساختاری موجود در شبکه گرافی استفاده می‌کند؛ بدین‌صورت که تمام این پیمایش‌ها به صورت رشته‌هایی به مدل زبانی Skip-Gram برای آموزش داده می‌شود. پیاده‌سازی این روش مبتنی بر مدل زبانی وردتووک است که در سال ۲۰۱۳ توسط گوگل معرفی شده است. روش DeepWalk با استفاده از راه‌ها و مسیرهای تصادفی موجود در گراف به شناسایی

الگوهای پنهان موجود در گراف می‌پردازد. الگوها یاد گرفته می‌شوند و توسط شبکه عصبی کدگذاری می‌شوند تا بردارهای تعبیه‌شده پایانی را تولید کنند. این مسیرهای تصادفی از یک ریشه هدف شروع می‌شوند و به صورت تصادفی بین همسایگان آن گره پیمایش می‌کنند و چندین گام پیش روند (با توجه به گام‌های اجرای الگوریتم).



شکل ۰-۵: الگوریتم DeepWalk



شکل ۰-۶: الگوریتم DeepWalk به صورت جزئی به همراه روش Skip-Gram که بخش از این الگوریتم است[۶۹]

تابع هدف الگوریتم DeepWalk به صورت زیر است:

$$min_y - logP(\{v_{i-w}, \dots, v_{i-1}, v_{i+1}, \dots, v_{i+w}\}|y_i) \quad (۶-۰)$$

که w اندازه پنجره برای محدود کردن طول گام‌های تصادفی است. SkipGram محدودیت‌های وجود را از بین می‌رود و فرمول ۳-۶ به صورت زیر تبدیل می‌شود:

$$min_y - log \sum_{-w \le j \le w} P(v_{i+j}|y_i) \quad (۷-۰)$$

در فرمول ۳-۷ $P(v_{i+j}|y_i)$ با استفاده از تابع سافت‌مکس تعریف می‌شود:

$$P(v_{i+j}|y_i) = \frac{\exp(y_{i+j}^T y_i)}{\sum_{k=1}^{|V|} \exp(y_k^T y_i)} \quad (۸-۰)$$

به دلیل پیچیدگی محاسبه جمع روی ضرب داخلی تمام گره‌های در گراف و هزینه محاسباتی زیاد، از

روش‌هایی برای حل این مشکل استفاده می‌شود مانند سافت‌مکس سلسله مراتبی و نمونه‌برداری منفی.

Algorithm 1. Framework of Generation of Item Embedding

Input: $Graph(V, E)$
Window size w //set $w = |S_i|$
Embedding size d //set d = item embedding dimensions
Walk per node γ // Number of random walks to start at each node
Walk length t // Length of the random walk started at each node
Output: node embedding, matrix of vertex represention $\phi \in R^{|V| \times d}$.

```
Initialize φ;
for i = 0 to φ do
    V' = Shuffle(V);
    for v_i in V' do
        nodeSequence = RandomWalk(G, v_i, t);//generate the nodesequences
        SkipGram(φ, nodeSequence, w); // generate node embedding
    end for
end for
```

شکل ۰-۷: شبه‌کد الگوریتم DeepWalk برای تعبیه گراف[۷۰]

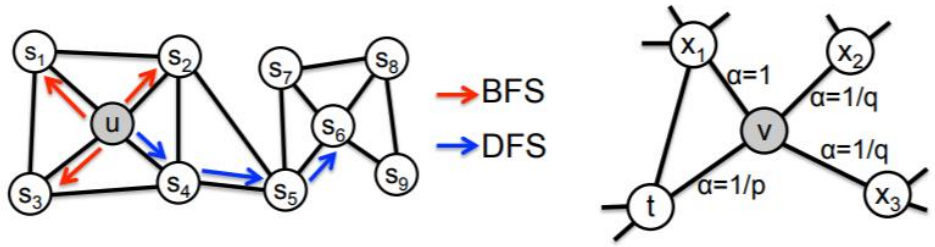
۲-۳-۳-۳ روش *Node2vec*

روش Node2vec نشان می‌دهد که روش DeepWalk به اندازه کافی در شناسایی الگوهای مختلف در شبکه قدرت ندارد و به تنهایی کافی نیست. این روش با استفاده از تعریف جدید برای همسایگی گره‌های شبکه و ایجاد استراتژی گام‌های تصادفی مرتبه دوم به نمونه‌برداری از گره‌ها می‌پردازد و از الگوریتم‌های جستجوی اولین سطحی[۱] و جستجوی اولین عمقی[۲] استفاده می‌کند. در واقع روش node2vec تفاوتی که با روش DeepWalk دارد این است که روش Node2vec از گام‌های تصادفی متعادل‌تری استفاده می‌کند و همانطور که گفته شد یک تعادلی بین روش‌های جستجوی سطحی گراف و جستجوی عمقی گراف برقرار می‌کند. که همین نیز باعث می‌شود این روش تعبیه نودهای با کیفیت بهتر و با اطلاعات بالاتری را ایجاد کند. برای ایجاد تعادل بین دو روش BFS و DFS از فراپارامترهایی p (نرخ بازگشت) و q (نرخ جستجو) استفاده می‌کنیم. فرض می‌شود که از گره t به گره v مسیری را طی کرده‌ایم و حال نیاز داریم که تصمیم بگیریم که گره‌ها x_i را جستجو کنیم. تعداد یال‌های بین t و گره‌های احتمالی x_i را محاسبه می‌کنیم. اگر فاصله صفر بوده (فقط t وجود دارد) وزن یال اصلی را با $\frac{1}{p}$ ضرب می‌کنیم. اگر فاصله یک

[1] Breadth-First Search
[2] Depth-First Search

بود، وزن یال اصلی را نگه می‌داریم و اگر فاصله ۲ بودن، وزن یال را در $\frac{1}{q}$ ضرب می‌کنیم. سپس مقدار وزن یال را برای دستیابی به مقدار احتمالاتی نرمال‌سازی می‌کنیم.

قابل توجه است که مقدار پایین $\frac{p}{q}$ منجر به احتمال جستجوی بالاتری می‌شود. به طور کلی در روشNode2vec از ساختارهای برابر و شباهت‌های گراف نیز در بازنمایی گره‌ها استفاده می‌شود.



شکل ۰-۸: الگوریتم Node2vec و روش‌های BFS و DFS. [۶۷]

Algorithm 1 The *node2vec* algorithm.

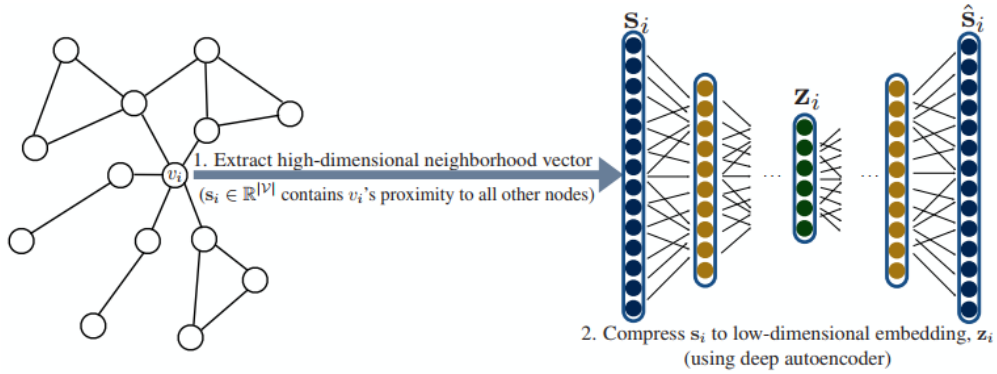
LearnFeatures (Graph $G = (V, E, W)$, Dimensions d, Walks per node r, Walk length l, Context size k, Return p, In-out q)
π = PreprocessModifiedWeights(G, p, q)
$G' = (V, E, \pi)$
Initialize $walks$ to Empty
for $iter = 1$ **to** r **do**
  for all nodes $u \in V$ **do**
    $walk$ = node2vecWalk(G', u, l)
    Append $walk$ to $walks$
f = StochasticGradientDescent(k, d, $walks$)
return f

node2vecWalk (Graph $G' = (V, E, \pi)$, Start node u, Length l)
Inititalize $walk$ to $[u]$
for $walk\_iter = 1$ **to** l **do**
  $curr = walk[-1]$
  V_{curr} = GetNeighbors($curr$, G')
  s = AliasSample(V_{curr}, π)
  Append s to $walk$
return $walk$

شکل ۰-۹: شبه‌کد الگوریتم Node2vec .[۷۰]

۳-۳-۳-۳ روش *SDNE*

در سال ۲۰۱۶ از ترکیب روش‌های شبکه خودرمزگزار و همجواری مرتبه اول و مرتبه دوم به معرفی این رویکرد پرداختند. بهینه‌سازی همزمان دو همجواری از ایده‌های اصلی این روش است. این روش از توابع غیرخطی برای رسیدن به تعبیه گراف و نودهای آن استفاده می‌کند. این مدل شامل دو روش بدون نظارت و با نظارت است. در این رویکرد برخلاف روش‌های تعبیه سطحی و غیرعمیق، از کل ساختار گراف استفاده می‌شود و ساختار کلی گراف به الگوریتم خودرمزگذار برای کدگذاری داده می‌شود.



شکل ۳-۱۰:شمای کلی تعبیه یک نود گراف با استفاده از رویکرد یادگیری عمیق شبکه‌های خودرمزگذار[۶۹]

روش SDNE خطای بازسازی داده‌ها ورودی و خروجی را کمینه می‌کند. تابع هزینه این روش در زیر آورده شده است:

$$l_{mix} = l_{2nd} + \alpha l_{1st} + \nu l_{regularization} \qquad (۳-۹)$$

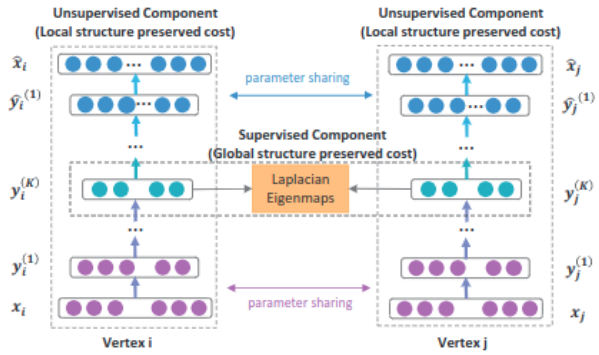
در تابع هزینه ۳-۹ مقدار α پارامتری است که توسط تابع هزینه همجواری مرتبه اول تنظیم می‌شود، $l_{regularization}$ شامل تنظیم نرم دوم برای کنترل آموزش بیش‌ازاندازه شبکه است و مقدار ν پارامتر برای تأثیر عبارت تنظیم است که استفاده شده در تابع هزینه.

(۳-۱۰)تابع هزینه همجواری مرتبه اول

$$\mathcal{L}_{1st} = \sum_{i,j=1}^{n} S_{i,j} \|y_i - y_j\|_2^2$$

(۳-۱۱)تابع هزینه همجواری مرتبه دوم

$$\mathcal{L}_{2nd} = \sum_{i=1}^{n} \|(\hat{x}_i - x_i) \odot w_i\|_2^2$$



شکل ۰-۱۱: الگوریتم SDNE و ساختار شبکه یادگیری عمیق درونی آن[۷۱]

جدول ۰-۵:الگوریتم‌های بازنمایی گراف و همجواری‌هایی ک از آن استفاده می‌کنند

الگوریتم	همجواری مرتبه اول	همجواری مرتبه دوم	همجواری مرتبه بالاتر
DeepWalk		●	●
Node2vec		●	●
SDNE	●	●	

۳-۳-۴ الگوریتم خوشه‌بندی

خوشه‌بندی از جمله روش‌هایی است که در آن هیچ‌گونه برچسبی برای رکوردها در نظر گرفته نمی‌شود و رکوردها فقط بر اساس معیار شباهتی که معرفی شده است، به مجموعه‌ای از خوشه‌های گروه‌بندی خواهند شد. عدم وجود برچسب سبب می‌شود که هر الگوریتم خوشه‌بندی یک الگوریتم بدون ناظر به حساب بیاید. در روش‌های بدون ناظر مراحلی تحت عنوان آموزش و ارزیابی وجود نداشته و در پایان عملیات خوشه‌بندی مدل ساخته شده به همراه کارایی براساس معیارهای مختلف در کاربردهای آن به عنوان خروجی داده می‌شود.

بعد از مرحله تعبیه گراف و به دست آوردن بردارهای ویژگی هر گره، در این مرحله با استفاده از الگوریتم یادگیری بدون نظارت به خوشه‌بندی و برچسب زدن داده‌ها می‌پردازیم. در این مرحله برای انجام خوشه‌بندی، در وهله‌ی اول انجمن‌های پرارزش به دست آمده را یکبار بدون بردار ویژگی به دست آمده از تعبیه گراف به همراه شاخص‌های ومعیارهایی که از گراف‌کاوی به دست آوردیم، الگوریتم

خوشه‌بندی می‌دهیم و در مقایسه با این کار، در مرحله‌ی بعد بردار ویژگی به دست آمده را به ویژگی‌های الگوریتم اضافه می‌کنیم و این بار خوشه‌بندی را با داده‌های جدید و تعداد خوشه‌های مشابه با مرحله قبل انجام می‌دهیم. انتظار داریم که خوشه‌بندی‌های انجام شده به همراه بردار ویژگی و شاخص‌ها و معیارهای گراف به دست آمده دارای کیفیت بیشتر و بهتری نسبت به حالت اول باشند. دلیل این امر نیز این است که در حالت اول از اطلاعات دیگر داده‌ها و گره‌هایی استفاده نمی‌شود، ولی بعد از اضافه کردن بردار ویژگی و شاخص‌ها و معیارهای گراف به دیگر ویژگی‌ها، اطلاعات گره‌های دیگر و همسایگان هر گره در خوشه‌بندی دخیل خواهد بود.

اساس کار الگوریتم‌های خوشه‌بندی میزان مشابهت داده‌ها است که براساس معیارهای مختلف اندازه‌گیری می‌شود که در هر حوزه و کاربردی معیار مختلفی مدنظر است. به طور کلی در همه‌ی معیارها، هدف کمینه کردن فاصله درون خوشه‌ای و بیشینه‌سازی فاصله میان خوشه‌هاست.

۳-۳-۴-۱ الگوریتم خوشه‌بندی $K-Means$

الگوریتم خوشه‌بندی K-Means یکی از ساده‌ترین و مشهورترین روش خوشه‌بندی بدون نظارت برای گروه‌بندی داده‌های بدون برچسب است. در این الگوریتم تعداد خوشه‌ها از قبل و به صورت ثابت تعریف شده است. این الگوریتم به صورت تکراری داده‌ها بدون برچسب را به خوشه‌های kتایی تقسیم می‌کند و با استفاده از معیارهای شباهت به خوشه‌بندی داده‌ها می‌پردازد. در این روش خوشه‌بندی داده‌های درون هر خوشه دارای بیشترین میزان مشابهت و خوشه‌های مختلف میزان شباهت کمتری دارند. هدف اصلی این الگوریتم (که مبتنی بر مرکز داده‌ها نیز است) کمینه ساختن مجموع فاصله‌های بین نقاط داده و خوشه متناظر است.

مراحل الگوریتم خوشه‌بندی $K-Means$:

۱. انتخاب تعداد خوشه‌ها که باید از ابتدا تعریف شود.

۲. انتخاب مراکز که به تعداد خوشه‌هایی است که در مرحله اول تعریف شد وابسته است.

۳. اختصاص هر داده به نزدیک‌ترین مرکز خوشه

۴. محاسبه میزان واریانس و تخصیص محل مرکز جدید به هر خوشه

۵. تکرار مرحله سوم

۶. اگر هیچ اختصاص جدیدی برای محل خوشه وجود نداشت الگوریتم خاتمه می‌یابد.

مشهورترین معیارهای فاصله که برای محاسبه فاصله بین داده‌ها وجود دارند معیارهای فاصله اقلیدسی و فاصله همینگ هستند که به ترتیب (۳-۱۲) و (۳-۱۳) در روابط ارائه شده‌اند.

$$d_E(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} \qquad (۱۲-۰)$$

$$d_H(x,y) = \sum_{k=1}^{n}|x_k - y_k| \qquad (۱۳-۰)$$

که در این دو رابطه به ترتیب n بیانگر ابعاد مسئله‌ی خوشه‌بندی و یا همان تعداد ویژگی‌هاست و همچنین x_k و y_k هم به ترتیب مبین kامین ویژگی‌های دور داده x و y هستند.

Algorithm 1 Iterative k-Means Algorithm

```
1: procedure ITERATIVE k-MEANS
2:     Select k points as initial centroids of the k clusters.
3:     for iterations := 1 to MAX_ITER do
4:         for each point p in the dataset do
5:             Assign point p to the cluster with the closest centroid
6:         end for
7:         for each cluster c do
8:             Recompute the centroid of c as the mean of all the data points assigned to c
9:         end for
10:    end for
11: end procedure
```

شکل ۰-۱۲: شبه‌کد الگوریتم خوشه‌بندی K-Means [۳۲]

۳-۳-۴-۲ تعیین تعداد خوشه‌ها با استفاده از الگوریتم Elbow Method

علاوه بر ویژگی‌های مثبتی که الگوریتم Kmeans دارد، یکی از مشکلاتی که این الگوریتم دارد، از تعریف کردن تعداد خوشه‌ها از قبل است. بدین معنی که باید تعداد خوشه‌ها را برای اجرای الگوریتم به

عنوان یکی از پارامترهای ورودی به الگوریتم خوشه‌بندی Kmeans بدهیم. از آنجایی که به دست آوردن تعداد دقیق خوشه‌ها در زمانی که داده‌ها برچسبی ندارد، با سعی و خطا به دست می‌آید، از این رو با استفاده از روش Elbow Method به شناسایی مقدار دقیق خوشه‌ها می‌پردازیم.

در تحلیل خوشه‌ها، این روش یک روش ابتکاری برای شناسایی تعداد خوشه‌هاست. این روش یک نمودار را در اختیار ما قرار می‌دهد که با استفاده از آن و مقادیر انحراف، تابع هزینه و یا خطا خوشه‌بندی می‌توان به تعداد دقیق خوشه‌ها به طور نسبی رسید. ایده اصلی این روش بدین صورت است که با توجه به افزایش تعداد خوشه‌ها، مقدار تطبیق داده‌ها به خوشه‌ها به طور طبیعی افزایش می‌یابد، به عنوان مثال خوشه اول بر اساس متغیرها مقدار زیادی اطلاعات را اضافه می‌کند، اطلاعات اضافه شده به خوشه اول، بیشترین مقدار تطبیق با خوشه‌بندی را دارد، با افزایش تعداد خوشه‌ها، این مقدار تطبیق بین داده‌ها کاهش می‌یابد. از آنجایی که پارامترهای مختلفی برای خوشه‌بندی وجود دارد، اما این تعداد خوشه‌ها می‌تواند از بیش‌برازش[1] داده‌ها جلوگیری کند.

در مرحله آخر بعد از انجام خوشه‌بندی و به دست آوردن معیارهایی در مورد کیفیت خوشه‌های به تحلیل در مورد خوشه‌ها و داده‌های موجود در آن می‌پردازیم و با استفاده از فرضیات موجود در مورد داده‌های ناهنجار هر خوشه یا خوشه‌هایی که رفتار متفاوتی از دیگر خوشه‌های دارند تصمیم می‌گیریم.

۳-۴ جمع‌بندی

در این فصل به ارائه رویکرد پیشنهادی و معماری اصلی پرداختیم. در ابتدای فصل به تعریف داده‌ها پرداختیم و در مرحله بعد، در مورد روش‌های تحلیل گراف و استخراج اطلاعات مفید از ساختار گرافی داده‌های پرداختیم. از گراف داده‌ها، انجمن‌های باارزش و مفید را استخراج کردیم. به دلیل پیچیدگی زیاد اجرای الگوریتم داده‌کاوی و یادگیری بر روی داده‌های گرافی و دقت نسبتاً پایین حاصل از اجرای این

[1] Overfitting

الگوریتم، با استفاده از تکنیک‌های تعبیه گراف، گره‌های گراف را تعبیه کردیم و بردار ویژگی‌ای از آن به دست آوردیم که اطلاعات دیگر گره‌ها را در آن وجود دارد. هدف اصلی پژوهش این است که تشخیص دهیم در حالت بدون استفاده از این بردار ویژگی و در حالت استفاده از این بردارهای ویژگی در کنار دیگر داده‌ها، به کیفیت خوشه‌های بهتری دست خواهیم یافت. و در مرحله آخر تحلیل بر روی خوشه‌ها و داده‌های هر خوشه پرداختیم.

فصل ۴ پیاده‌سازی، آزمایش‌ها و ارزیابی

۴-۱ مقدمه

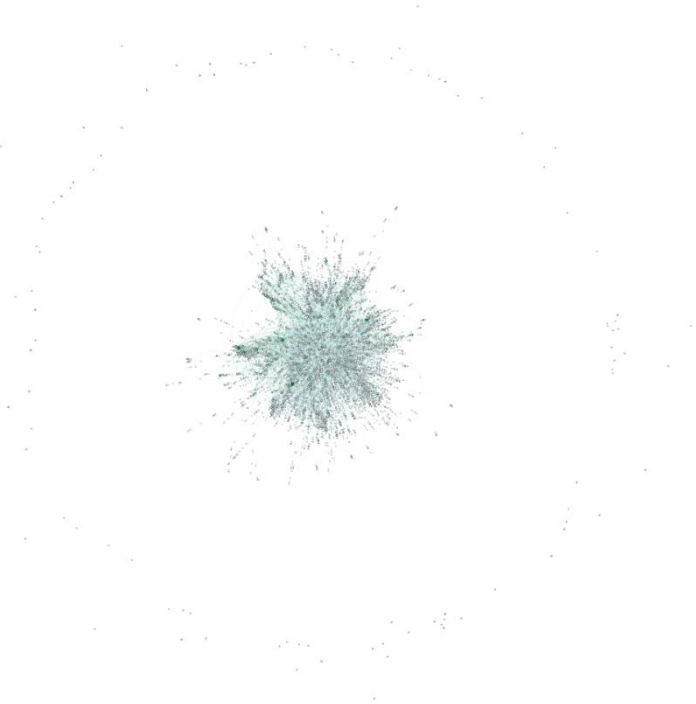
با توجه به شرح کامل رویکرد پیشنهادی در فصل سوم، در این فصل به بررسی آزمایش‌های دقیق و جامع و پیاده‌سازی‌های مربوطه می‌پردازیم. همانطور که گفته شد، روش پیشنهادی شامل سه بخش است، در بخش اول داده‌های مسئله را به مدل گرافی تبدیل می‌کنیم و معاملات را در قالب گرافی تحلیل می‌کنیم سپ از گراف معاملاتی اطلاعات مفید را توسط معیارهای مختلف استخراج می‌کنیم و انجمن‌های باارزش را جدا می‌کنیم. دلیل جدا کردن انجمن‌های باارزش سربار محاسباتی زیاد و معاملات با تعداد کم ارزش ریالی معاملات بود که در این بخش انجام شد. در بخش دوم با اعمال روش‌های تعبیه گراف (که در اینجا تعبیه گره‌ها مدنظر است) بر روی این انجمن‌های باارزش، ویژگی‌های تعبیه‌شده برای هر گره را به دست می‌آوریم و در مرحله آخر با استفاده از الگوریتم‌های خوشه‌بندی به خوشه‌بندی و تحلیل و استخراج ویژگی‌های هر خوشه و داده‌ها که رفتار متفاوت با دیگر داده‌ها دارند می‌پردازیم. در ادامه فصل ابتدا به تعریف مجموعه دادگان می‌پردازیم.

۴-۲ شناسایی انجمن‌های باارزش و تحلیل گراف‌های آن‌ها

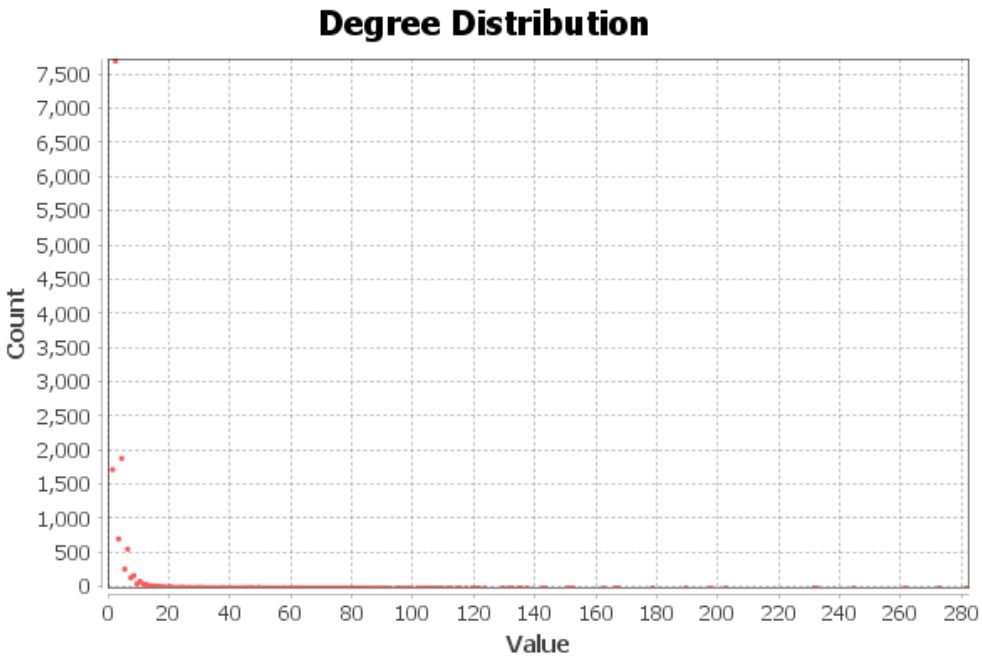
پس از معرفی مجموعه دادگان، با استفاده از نرم‌افزار Gephi به ایجاد ساختار گرافی اطلاعات پرداختیم. این نرم‌افزار که در سال ۲۰۰۹ ایجاد شده به صورت متن‌باز است و برای تحلیل شبکه‌های اجتماعی و بصری سازی اطلاعات در قالب گراف به کار می‌رود و در بستر زبان برنامه‌نویسی جاوا نوشته شده است. از ویژگی‌های مهم این نرم‌افزار رایگان بودن، سادگی کار با نرم‌افزار و قابلیت‌های بسیار برای تحلیل شبکه‌های اجتماعی و انجام تئوری‌ها بر روی گراف است که اطلاعات بسیاری را در اختیار کاربر قرار می‌دهد. گراف اطلاعاتی به صورت معاملات انجام شده بین مناقصه‌گزاران و مناقصه‌گران با یال بدون جهت است و برای هر گره ویژگی‌هایی را در نظر گرفتیم. در مرحله پیش‌پردازش بر روی گراف ، دستگاه‌های مناقصه‌گزاری که فقط یک معامله داشتند را از ساختار گراف حذف کردیم و سپس با تحلیل گراف و شناسایی انجمن در گراف به دو شبکه گرافی در هر گراف اطلاعاتی (گراف با وزن تعداد معاملات

و گراف با وزن ارزش ریالی معاملات) دست یافتیم. سپس شاخص‌های هر شبکه گرافی را تحلیل کرده‌ایم.

همانطور که گفته شد در این پژوهش دید ما نسبت به گراف به دو صورت بوده است. گراف با وزن یال تعداد معاملات و گراف با وزن ارزش ریالی معاملات. در مرحله‌ی بعد به تحلیل دو شبکه گراف بین مناقصه گران و مناقصه‌گزاران پرداخته شد، گراف بدون جهت با وزن یال تعداد معامله و شبکه گرافی بدون جهت با وزن ارزش ریالی معاملات. شاخص‌ها برای هرکدام از شبکه‌ها در جدول نشان داده شده است.

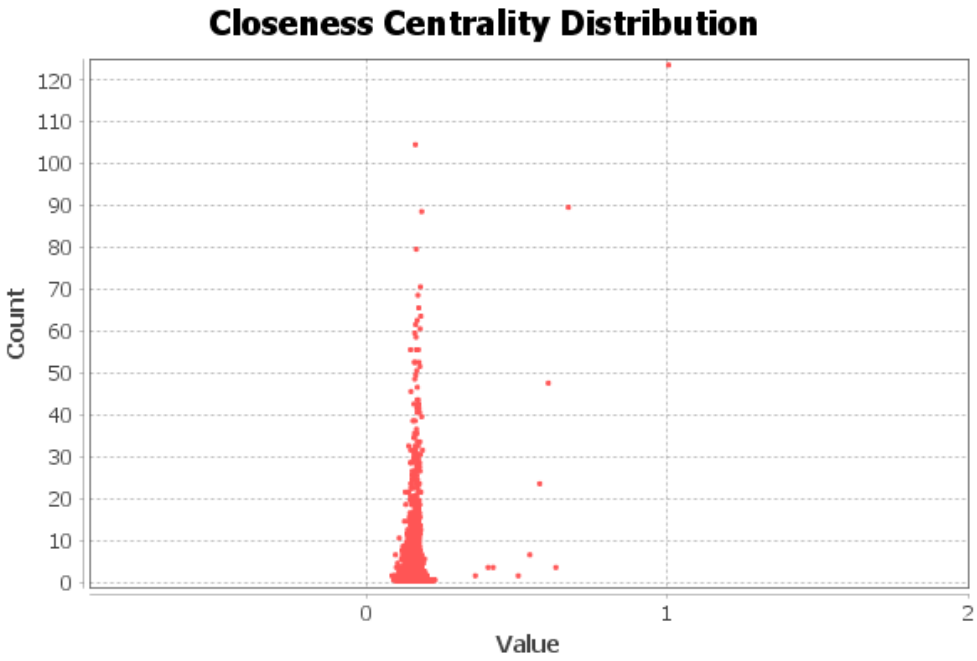


شکل ۰-۱:گراف کلی معاملات بین مناقصه‌گزاران و مناقصه گران که شامل ۱۴۲۰۶ گره و ۱۹۹۰۸ یال است



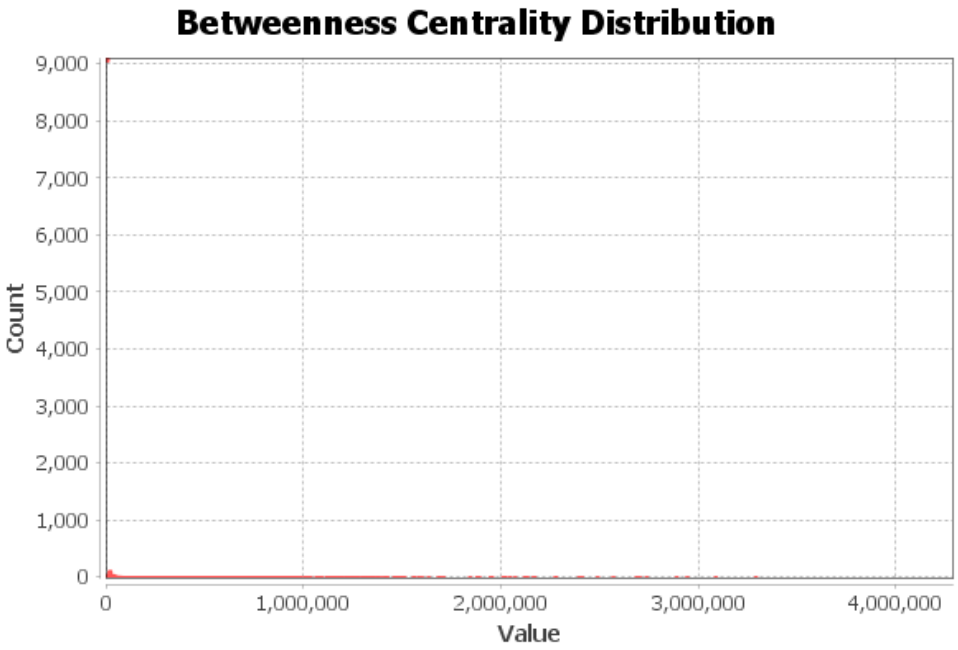
شکل ۲-۲۰:پراکندگی وزن گره‌ها در گراف معاملات

میانگین درجه برای گره‌های از نوع دستگاه اجرایی بیانگر متوسط تعداد تأمین‌کنندگان در معاملات دستگاه اجرایی است و برای گره‌هایی از نوع تأمین‌کننده بیانگر متوسط تعداد دستگاه اجرایی است که تأمین‌کننده با آن در ارتباط بوده است. پراکندگی وزن درجه‌ها در شکل ۲ نشان داده شده است. با توجه به شکل ۴-۲ می‌توان دریافت که تعداد نسبتاً زیادی از گره‌ها درجه یک دارند. در گراف کلی معاملات هر مناقصه‌گزار حداقل با یک و حداکثر با ۱۵۸ تأمین‌کننده در ارتباط بوده و به طور میانگین با ۱۴,۲ تأمین‌کننده تعاملاتی داشته است. هر تأمین‌کننده نیز حداقل با یک و حداکثر با ۳۷ مناقصه گزار در ارتباط بوده و به طور میانگین با ۱,۵۵ مناقصه گزار تعاملاتی داشته است.



شکل ۳-۰: شاخص مرکزیت نزدیکی گره‌ها در گراف معاملات

همانطور که گفته شد، شاخص مرکزیت نزدیکی معیاری برای هر گره است که میزان نزدیکی به دیگر نودها را اندازه‌گیری می‌کند. بدین معنی که هر گره چه تعداد همسایه دارد و با همسایگان خود چه میزان نزدیک است. با توجه به شکل ۳-۴، شاخص مرکزیت نزدیک مقداری بین صفر و یک است و میزان نزدیکی داده‌ها به هم در بیشتر گره‌ها نزدیک به مقدار میانگین ۰,۱۷ است.



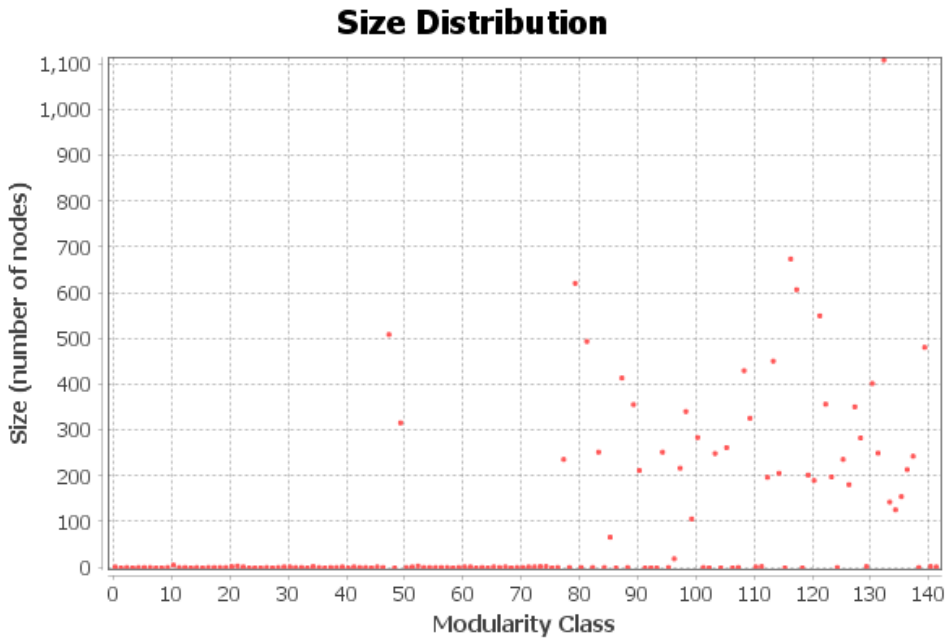
شکل ۰-۴: شاخص مرکزیت بینابینی در گراف معاملات

همانطور که گفته شد شاخص مرکزیت بینابینی بیانگر جایگاه یک گره در شبکه است. هرچه گره‌های شبکه برای ایجاد ارتباط با سایر گره‌ها به یکدیگر وابسته باشند، آن گره قدرت بیشتری در شبکه خواهد داشت. میانگین شاخص مرکزیت بینابینی در هر دو شبکه ۰,۰۰۰۴ محاسبه شده است.

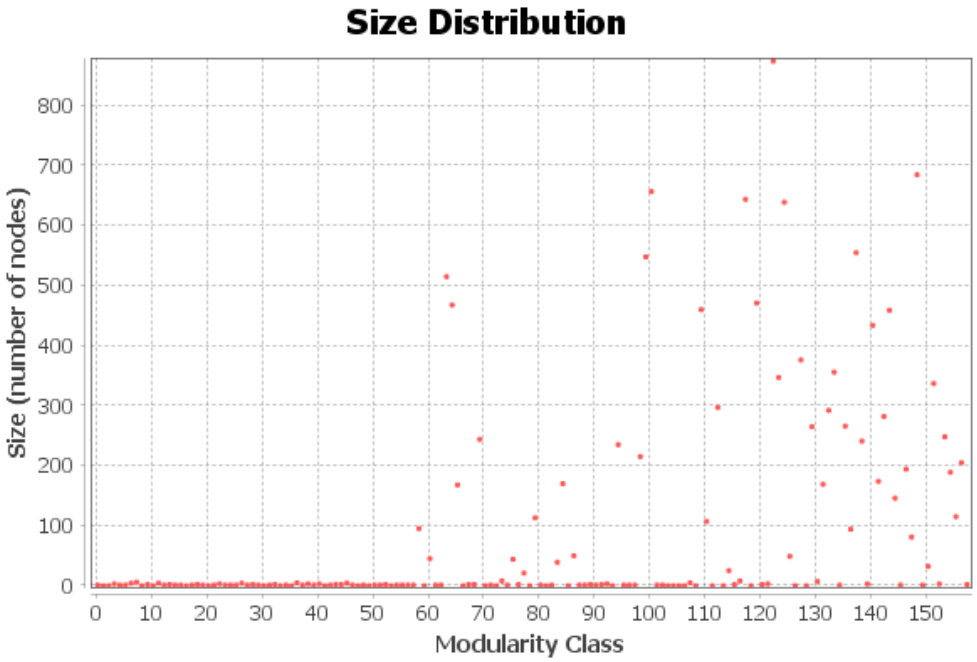
جدول ۰-۱: شاخص‌ها و سنجه‌های محاسبه شده برای شبکه

نوع گره	تعداد گره	فراوانی	میانگین درجه	میانگین مرکزیت بینابینی	میانگین مرکزیت نزدیکی	مرکزیت بینابینی	قدرت
مناقصه گزار	۱۳۹۸	۹,۸	۱۴,۲	۰,۰۰۲۱	۰,۲۳	۰,۰۰۳۹	۰,۰۰۱۱
مناقصه گر(تأمین‌کننده)	۱۲۸۰۸	۹۰,۲	۱,۵۵	۰,۰۰۰۲	۰,۱۷	۰,۰۰۰۴	۰,۰۰۱۵

یکی از مهم‌ترین ویژگی‌های هر شبکه گرافی تشخیص وابستگی‌ها و شناسایی انجمن‌ها است. یکی از الگوریتم‌هایی که برای شناسایی انجمن‌ها به کار می‌رود، ماژولاریتی است. در شبکه گرافی بدون جهت با وزن تعداد معاملات ۱۴۲ انجمن با شاخص ماژولاریتی ۰,۸۵ شناسایی شده است و در گراف بدون جهت با وزن ارزش ریالی معاملات ۱۵۸ انجمن شناسایی شده که شاخص ماژولاریتی آن ۰,۸۹ است.



شکل ۰-۵: توزیع اندازه ماژول‌ها در انجمن‌های شناسایی شده در گراف بدون جهت با وزن تعداد معاملات

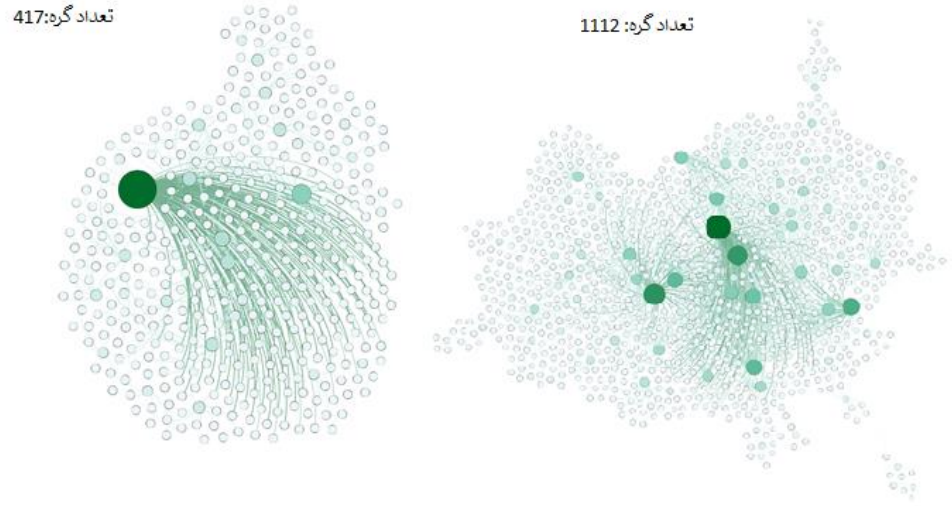


شکل ۰-۶:توزیع اندازه ماژول‌ها در انجمن‌های شناسایی شده در گراف بدون جهت با وزن ارزش ریالی معاملات

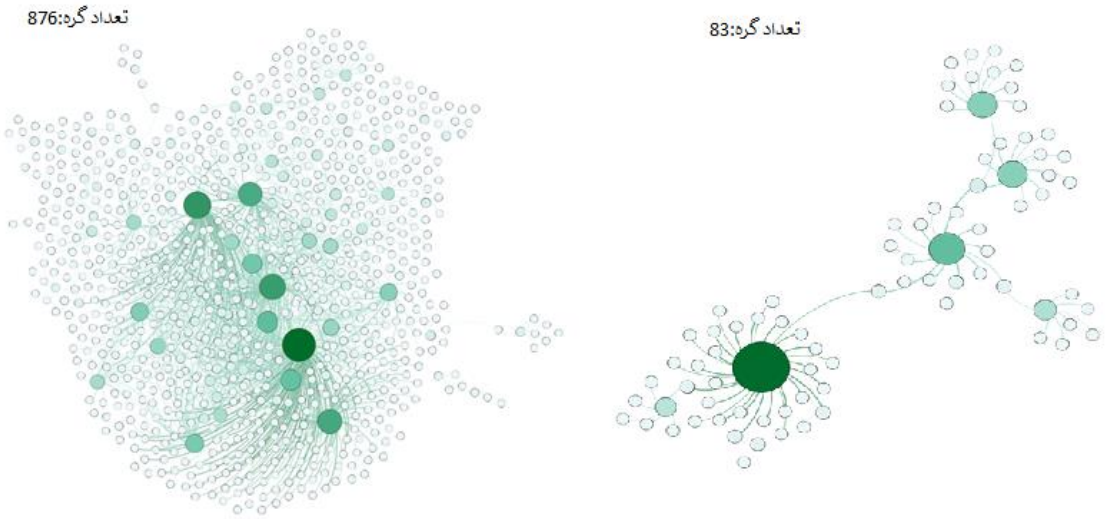
بعد از تحلیل ماژولاریتی برای هر دو شبکه گرافی به تحلیل انجمن‌های باارزش و شناسایی آن‌ها می‌پردازیم. در جدول ۴-۲ بزرگ‌ترین‌ها انجمن‌های شناسایی شده در گراف بدون جهت با وزن تعداد معاملات و گراف بدون جهت با وزن ارزش ریالی معاملات آورده شده است.

جدول ۰-۲: بزرگ‌ترین و مؤثرترین انجمن‌های استخراج شده از گراف معاملات کلی

نوع گراف	تعداد گره	تعداد یال	فراوانی
گراف بدون جهت با وزن یال تعداد معاملات	۱۱۱۲ گره	۱۶۴۹	مناقصه‌گزار: ۱۰,۷درصد تأمین‌کننده: ۸۹,۳درصد
	۴۱۷	۴۹۳	مناقصه‌گزار: ۸,۴درصد تأمین‌کننده: ۹۱,۶درصد
گراف بدون جهت با وزن ارزش ریالی معاملات	۸۷۶	۱۳۸۹	مناقصه‌گزار: ۱۰,۴درصد تأمین‌کننده: ۸۹,۶درصد
	۸۳	۸۲	مناقصه‌گزار: ۱۰,۸درصد تأمین‌کننده: ۸۹,۲درصد



شکل ۰-۷: انجمن‌های بررسی شده در گراف بدون جهت با وزن تعداد معاملات



شکل ۰-۸:انجمن‌های بررسی شده در گراف بدون جهت با وز ارزش ریالی معاملات

۴-۳ تعبیه گراف و کاهش ابعاد

همانطور که در فصل سوم توضیح دادیم، در این مرحله بعد از تحلیل گراف و به دست آوردن انجمن‌های باارزش برای اجرای بهتر و بازنمایی گراف باید گره‌های گراف را در ابعاد دیگری نمایش دهیم و برداری از ویژگی‌ها را به دست بیاوریم که اطلاعات گراف را در اختیار داشته باشد. برای تعبیه گراف از الگوریتم‌های گام‌های تصادفی (DeepWalk و Node2vec) و روش‌های یادگیری عمیق (SDNE) استفاده کردیم. در جدول زیر ابتدا پارامترهای هر یک از الگوریتم‌ها که را تنظیم کرده‌ایم مشخص می‌کنیم و خروجی این مرحله که گره‌های تعبیه‌شده گراف هستند را به عنوان ویژگی‌های اضافی به ویژگی‌های گراف اصلی که شامل تعداد معاملات و ارزش ریالی معاملات است در هر دو شبکه گرافی اضافه می‌کنیم و مجموع داده‌های اصلی به همراه این اطلاعات تعبیه‌شده گراف را به الگوریتم یادگیری ماشین و داده‌کاوی برای انجام خوشه‌بندی می‌دهیم و داده‌های هر خوشه را تحلیل می‌کنیم. برای تعبیه گراف و بازنمایی گراف و به دست آوردن بردار ویژگی برای هر گره ابتدا باید پارامترهای الگوریتم بازنمایی گراف را مشخص کنیم. برای اجرای هر یک از الگوریتم‌های بازنمایی گراف باید ورودی، خروجی، نوع گراف که به عنوان ورودی به الگوریتم داده می‌شود، ابعاد بازنمایی گراف که تعداد ابعاد پنهان برای

یادگیری هر گره است، جهت‌دار بودن یا نبودن گراف و وزن گراف است.

- ورودی الگوریتم← فایل ورودی گراف است که به صورت فایل متنی است و شامل دو ستون است و گره مبدأ و مقصد و یال‌های آن‌هاست. در واقع ورودی الگوریتم، گرافی شامل گره و یال‌هاست.
- نوع گراف ورودی← که به صورت لیستی از یال‌های یا ماتریس مجاورت است که در اینجا به صورت لیست یال (edgelist) است.
- نوع الگوریتم بازنمایی گراف←که شامل الگوریتم‌های DeepWalk، Node2vec، SDNE و دیگر روش‌هاست.
- اندازه بازنمایی گراف←ابعاد بردار ویژگی که برای یادگیری هر گره ایجاد می‌شود که می‌تواند هر مقداری باشد.
- اگر گراف وزن‌دار و یا جهت‌دار باشد نیز ویژگی‌ها باید جداگانه به صورت فایل به الگوریتم داده شود.
- خروجی الگوریتم← فایل متنی که شامل تعداد گره و بردار ویژگی هر گره که با ابعاد مشخص شده ایجاد می‌شود.

جدول ۰-۳: الگوریتم‌های بازنمایی گراف و پارامترهای آن

الگوریتم بازنمایی گراف	پارامترهای الگوریتم و مقادیر پیش‌فرض آن‌ها				
DeepWalk	تعداد گام‌ها که به صورت تصادفی از هر گره شروع می‌شود: **۱۰**	طول هر گام تصادفی که از هر گره شروع می‌شود: **۸۰**	اندازه پنجره Skip-Gram: **۱۰**		
Node2vec	مقادیر $\mathbf{p = 1.0}$ و $\mathbf{q = 1.0}$				
SDNE	لیست انکودر که لیستی شامل تعداد نورون‌های در لایه انکودر است و دومین مقدار لیست ابعاد خروجی اندازه بازنمایی دانش را نشان می‌دهد $[\mathbf{128} \quad \mathbf{1000}]$	مقدار α که فراپارامتری برای کنترل مقدار تابع هزینه همجواری مرتبه اول است: $\mathbf{1 * 10^{-6}}$	مقدار β که برای ساخت ماتریس B استفاده می‌شود: **۵**	مقدار $\mu_1 = \mathbf{1 * 10^{-5}}$ و $\mu_2 = \mathbf{1 * 10^{-4}}$ که برای کنترل مقادیر تابع هزینه l_1 و l_2 استفاده می‌شود.	نرخ یادگیری $\boldsymbol{learning\ rate = 0.01}$

در جدول زیر زمان اجرای الگوریتم برای هر یک از روش‌های و بر روی هر یک از گراف‌ها و انجمن‌های باارزشی که در قسمت قبل به دست آورده‌ایم را نمایش می‌دهیم.

جدول ۴-۰: زمان اجرای الگوریتم‌های تعبیه گراف (برحسب میلی‌ثانیه و ثانیه)

		$DeepWalk$(میلی‌ثانیه)		$Node2Vec$(میلی‌ثانیه)		$SDNE$(ثانیه)	
ابعاد بردار ویژگی		$d = 64$	$d = 128$	$d = 64$	$d = 128$	$d = 64$	$d = 128$
گراف بدون جهت با وزن تعداد معاملات	گره:۴۱۷ یال:۴۹۳	10.28	12.80	9.46	10.34	2.48	3.88
	گره:۱۱۱۲ یال:۱۶۴۹	33.50	41.38	28.09	30.76	4.90	4.96
گراف بدون جهت با وزن ارزش ریالی معاملات	گره:۸۷۶ یال:۱۳۸۹	24.65	30.57	21.84	23.71	3.86	3.99
	گره۸۳ یال:۸۲	1.08	1.29	1.41	1.57	1..88	1.96

همانطور که در جدول ۴-۵ مشاهده می‌کنیم زمان اجرای الگوریتم‌های تعبیه گراف با توجه به ابعاد بردار ویژگی تغییر می‌کند، و برای ابعاد بالاتر مدت زمان بیشتری برای به دست آوردن تعبیه گراف و ویژگی‌های هر گره و مشابهت آن با دیگر گره‌ها از گراف نیاز است. لازم به ذکر است که با تغییر پارامترهای هر یک از الگوریتم‌ها نیز این مدت زمان تغییر می‌کند. به عنوان مثال در الگوریتم $DeepWalk$ و در ابعاد ۱۲۸ با تغییر طول گام از ۸۰ به ۱۰۰ مدت زمان اجرای الگوریتم به 16.17 میلی‌ثانیه و با تغییر تعداد گام‌های تصادفی از پیش‌فرض ۱۰ به مقدار ۲۰ مدت زمان اجرای الگوریتم به 25.01 میلی‌ثانیه می‌رسد.

نکته‌ی قابل توجه مدت زمان نسبتاً بالای الگوریتم SDNE است که همانطور که در فصل سوم گفته شد یکی از مشکلات الگوریتم‌های یادگیری عمیق هزینه زمانی نسبتاً بالای آن‌هاست که دلیل این امر این است که علاوه بر اینکه ساختار محلی را در نظر می‌گیرد، با استفاده از همجواری مرتبه دوم ساختار کلی

گراف را در نظر می‌گیرد.

۴-۴ الگوریتم خوشه‌بندی $KMeans$

بعد از انجام تعبیه گراف و به دست آوردن بردارهای ویژگی برای هر گره از گراف‌ها، نوبت به مرحله‌ی انجام خوشه‌بندی و تحلیل بیشتر بر روی داده‌ها رسیده است. برای انجام خوشه‌بندی در دو مرحله عمل می‌کنیم. در وهله‌ی اول خوشه‌بندی را بر روی ویژگی‌های گره‌ها به همراه شاخص‌ها و معیارهای گراف (درجه، مرکزیت نزدیکی و مرکزیت بینابینی) که در مرحله اول به دست آوردیم، بدون اضافه کردن بردار ویژگی حاصل از مرحله تعبیه گراف انجام می‌دهیم و در مرحله‌ی بعد با اضافه کردن بردار ویژگی هر گره به دیگر ویژگی‌ها، الگوریتم خوشه‌بندی را انجام می‌دهیم.

جدول ۰-۵:پارامترهای الگوریتم خوشه‌بندی

تعداد دفعاتی که الگوریتم اجرا می‌شود	تعداد خوشه‌ها	تعداد تکرار	مقداردهی اولیه	Algorithm	پارامترها
10	Elbow method	300	k-means++	‘auto’	مقادیر

۴-۵ معیار ارزیابی

برای ارزیابی خوشه‌های به دست آمده از طریق الگوریتم خوشه‌بندی K-Means نیاز به روشی است که با توجه به نداشتن برچسب داده‌ها از پیش، معیار درستی از کیفیت خوشه‌های ایجاد شده را در اختیار ما بگذارد. همانطور که گفته شد معیار خوشه‌بندی درست و ایجاد خوشه‌هایی که به خوبی از هم جدا شده باید به گونه‌ای باشد که فاصله‌ی میان داده‌های درون هر خوشه کم و فاصله‌ی خوشه‌ها از هم زیاد باشد. معیاری که برای تعیین کیفیت خوشه‌ها در این پژوهش استفاده شده است، معیار Silhouette Coefficient است. دلیل استفاده از این معیار این است که به دلیل اینکه خوشه‌ها اندازه‌ها متفاوتی دارند، این معیار برای چنین مسئله‌ای عملکرد بهتری دارد. این معیار بیانگر میزان شباهت داده‌های درون خوشه

به یکدیگر (پیوستگی بین داده‌های یک خوشه) نسبت به دیگر خوشه‌ها و فاصله‌ی دیگر خوشه‌هاست. مقدار این معیار از یک تا منفی یک تغییر می‌کند که مقدار ۱ نشان‌دهنده‌ی این است که داده با مقدار مشابهت بالایی درون یک خوشه قرار دارد و از خوشه‌های دیگر و همسایه‌های خود دورتر است و شباهتی به خوشه‌های دیگر ندارد.

برای تعریف این معیار ابتدا فرض می‌کنیم که داده‌ها با هر روش خوشه‌بندی‌ای در تعداد خوشه‌های از پیش تعیین شده خوشه‌بندی شده‌اند. برای هر داده $i \in C_I$ که مقدار هر داده i در خوشه C_I است، داریم:

$$a(i) = \frac{1}{|C_I| - 1} \sum_{i \in C_I, i \neq j} d(i,j) \qquad (۱-۰)$$

که میانگین فاصله بین i و همه دیگر داده‌ها در همان خوشه است و $|C_I|$ تعداد داده‌هایی هست که به خوشه‌ی i متعلق است و $d(i,j)$ فاصله بین داده i و j در خوشه‌ی $|C_I|$ است. به طور کلی مقدار $a(i)$ مقداری است که نشان می‌دهد که چقدر i به یک خوشه به طور مناسبی متعلق است. سپس باید میانگین شباهت بین داده i و خوشه‌ی C_j به صورت میانگین فاصله‌ی بین i و خوشه‌های C_j است و داریم:

$$b(i) = \min_{J \neq I} \frac{1}{|C_I|} \sum_{i \in C_j} d(i,j) \qquad (۲-۰)$$

مقدار $b(i)$ کوچک‌ترین میانگین فاصله بین داده i و دیگر خوشه‌هاست که i عضو آن خوشه نیست. سپس با استفاده از مقادیر $a(i)$ و $b(i)$ به محاسبه‌ی مقدار معیار Silhouette می‌پردازیم:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, if |C_I| > 1 \qquad (۳-۰)$$

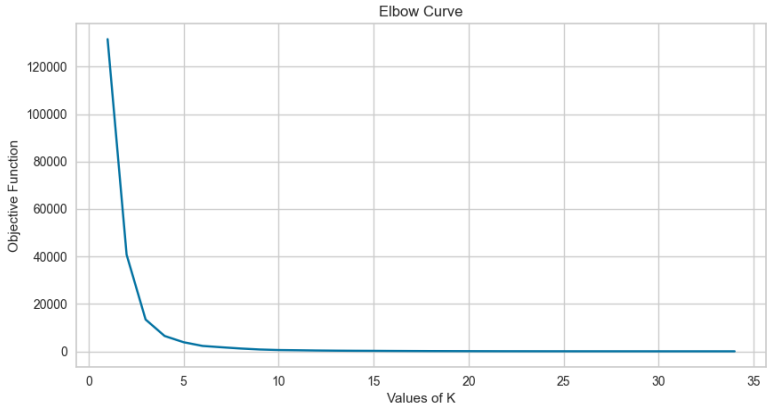
که داریم:

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)}, if a(i) < b(i) \\ 0 \quad\quad , if\ a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1, \quad if a(i) > b(i) \end{cases} \qquad (۴-۰)$$

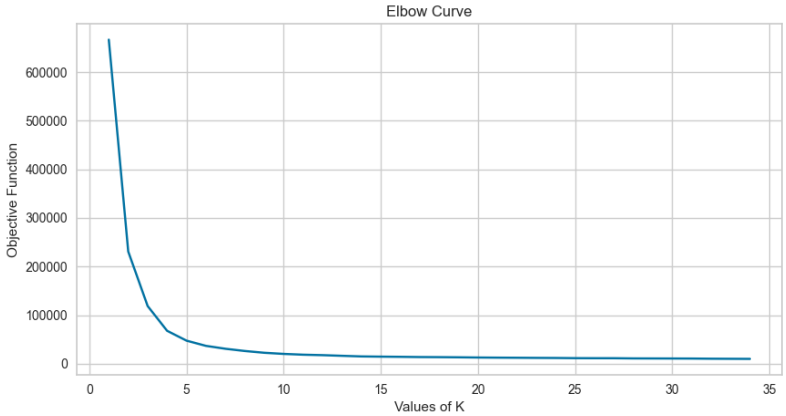
و مقدار $-1 < s(i) < 1$ است.

۴-۶ نتایج

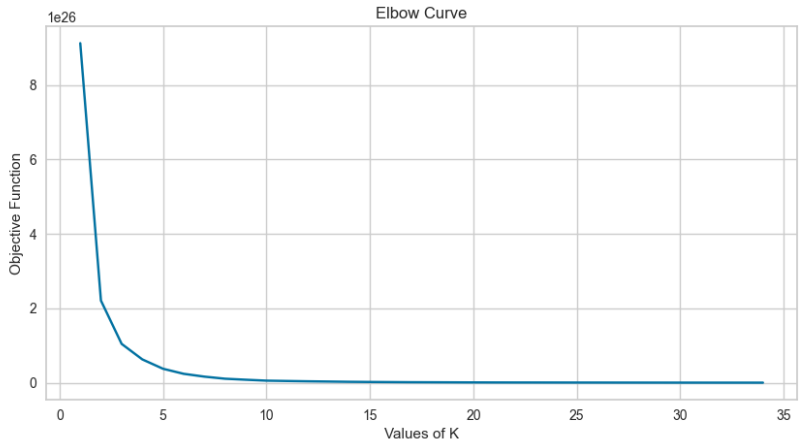
ابتدا همانطور که پیش‌تر گفته شد باید از طریق Elbow Method تعداد خوشه‌ها را تعیین کنیم. در مرحله بعد از شناسایی تعداد خوشه‌ها به خوشه‌بندی با استفاده از الگوریتم خوشه‌بندی KMeans می‌پردازیم و از معیار Silhoutte برای کیفیت خوشه‌ها استفاده می‌کنیم. در مرحله‌ی بعد اطلاعات هر خوشه را برای شناسایی و رفتار داده‌ها مورد بررسی قرار می‌دهیم و در جداول جداگانه به بررسی اطلاعات در مورد هر خوشه می‌پردازیم.



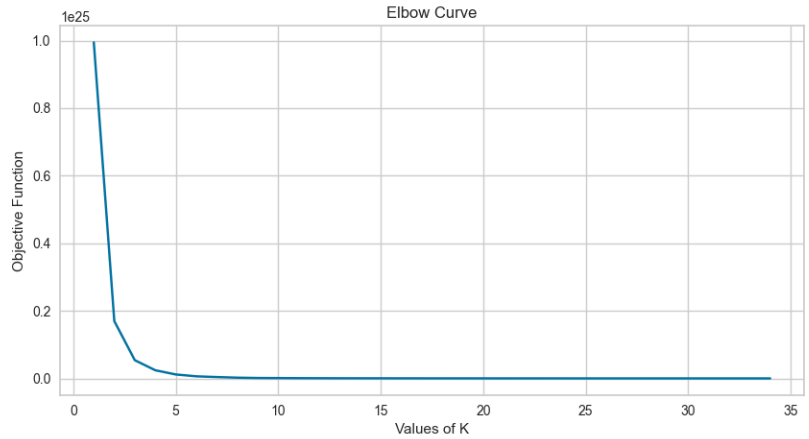
شکل ۴-۹: تعداد خوشه‌ها برای گراف بدون جهت با وزن تعداد معاملات با ۴۱۷ گره (که مقدار ۵ بهترین مقدار است)



شکل ۴-۱۰:تعداد خوشه‌ها برای گراف بدون جهت با وزن تعداد معاملات با ۱۱۲ گره (که مقدار ۴ بهترین مقدار است)



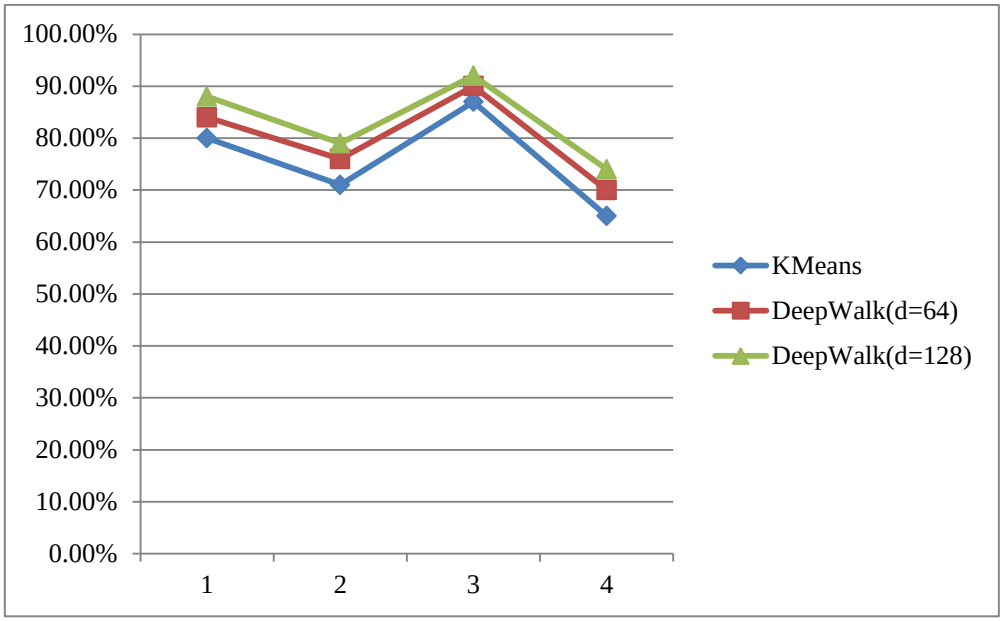
شکل ۰-۱۱: تعداد خوشه‌ها برای گراف بدون جهت با وزن ارزش ریالی معاملات با ۸۷۶ گره(که بهترین مقدار برای خوشه‌ها ۴ است)



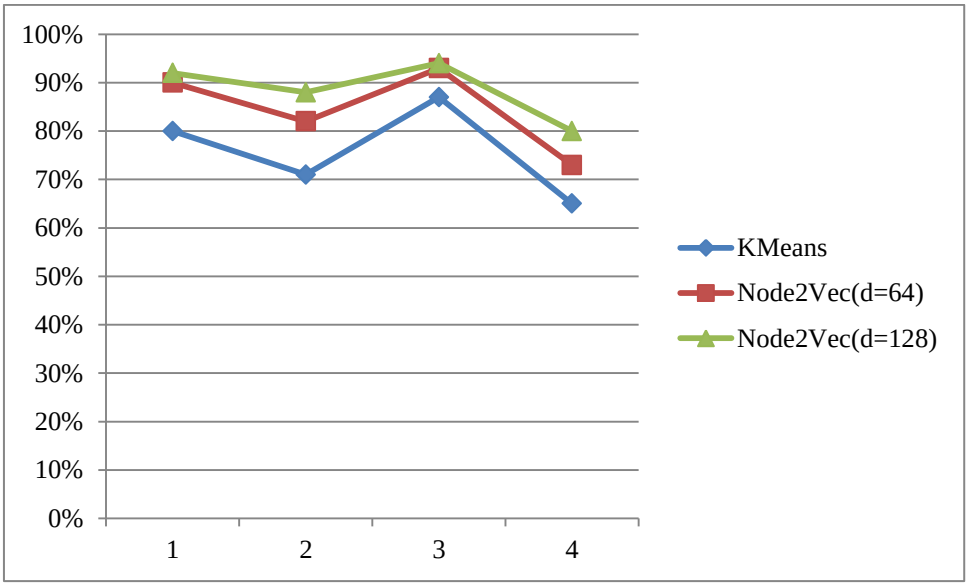
شکل ۰-۱۲:تعداد خوشه‌ها برای گراف بدون جهت با وزن ارزش ریالی معاملات با ۸۳ گره(که بهترین مقدار برای خوشه‌ها ۳ است)

جدول ۰-۶: مقایسه معیار خوشه‌بندی قبل و بعد از تعبیه گراف و اضافه کردن ویژگی‌ها بر اساس میانگین معیار Silhoutte

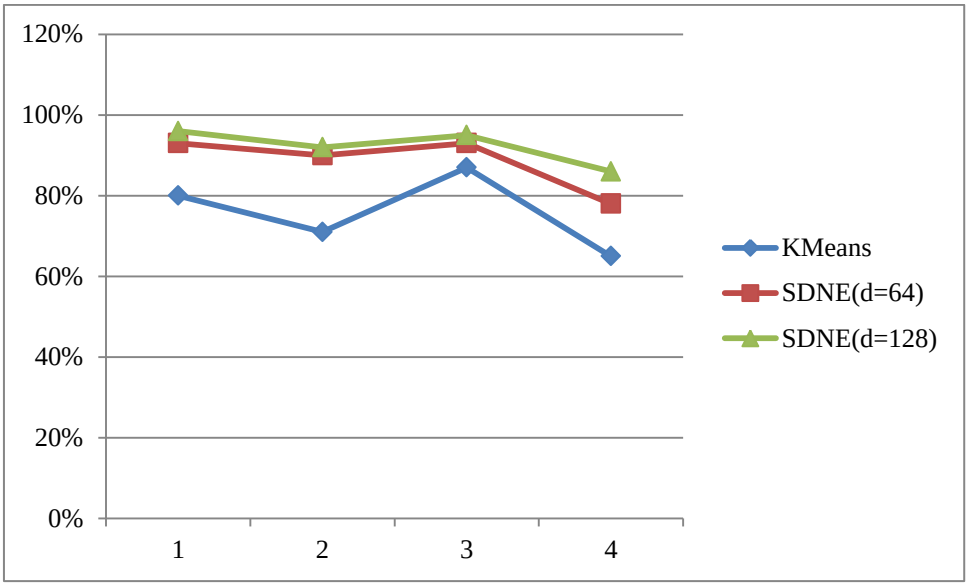
اطلاعات گراف		$KMeans - Silhoutte$	$DeepWalk$		$Node2Vec$		$SDNE$	
			$d = 64$	$d = 128$	$d = 64$	$d = 128$	$d = 64$	$d = 128$
گراف بدون جهت با وزن تعداد معاملات	گره: ۴۱۷ یال: ۴۹۳	0.80	0.84	**0.86**	0.90	**0.92**	0.93	**0.96**
	گره: ۱۱۱۲ یال: ۱۶۴۹	0.71	0.76	**0.79**	0.82	**0.88**	0.90	**0.92**
گراف بدون جهت با وزن ارزش ریالی معاملات	گره: ۸۷۶ یال: ۱۳۸۹	0.87	0.90	**0.92**	0.93	**0.94**	0.93	**0.95**
	گره۸۳ یال:۸۲	0.65	0.70	**0.74**	0.73	**0.80**	0.78	**0.86**



شکل ۰-۱۳: مقایسه خوشه‌بندی براساس معیار Silhoutte با استفاده از تعبیه گراف به همراه دیگر ویژگی‌ها و بدون تعبیه گراف در هر چهار گراف و در روش DeepWalk



شکل ۰-۱۴: مقایسه خوشه‌بندی براساس معیار Silhoutte با استفاده از تعبیه گراف به همراه دیگر ویژگی‌ها و بدون تعبیه گراف در هر چهار گراف و در روش node2vec



شکل ۰-۱۵: مقایسه خوشه‌بندی براساس معیار Silhoutte با استفاده از تعبیه گراف به همراه دیگر ویژگی‌ها و بدون تعبیه گراف در هر چهار گراف و در روش SDNE

با توجه به جدول ۴-۶ می‌توان نتیجه گرفت به دلیل اینکه در دفعه‌ی اول فقط بر روی داده‌های هر گراف خوشه‌بندی انجام شد و دیگر ویژگی‌های گراف و اطلاعات همسایگان آن‌ها در نظر گرفته نشد، مقدار میانگین خوشه‌بندی نتایج پایین‌تری را نسبت به مرحله بعد که اضافه کردن ویژگی‌های حاصل از تعبیه گراف است دارد. دلیل این امر این است که زمانی که برای خوشه‌بندی از اطلاعات دیگر گره‌ها استفاده می‌کنیم، مشابهت و خوشه‌بندی دقت بالاتری خواهد داشت. به عنوان زمانی که در الگوریتم DeepWalk طول گام تصادفی را مقدار پیش‌فرض ده در نظر می‌گیرم، بدین معنی است که به طور تصادفی از گره گراف شروع می‌کند و اطلاعات ده گره دیگر را نیز در خود دارد، که این ویژگی‌ها در قالب بردار ویژگی به ویژگی‌های دیگر گراف اضافه می‌شود و باعث دقت بالاتر و کیفیت بهتر خوشه‌ها می‌شود.

با توجه اشکال و نمودارها، می‌توان دریافت که اگرچه مدت زمان اجرای الگوریتم SDNE به میزان قابل توجهی بالاست، ولی به دلیل اینکه از همجواری مرتبه اول و همجواری مرتبه دوم برای به دست آوردن بردار ویژگی استفاده می‌کند و علاوه بر ساختارهای محلی، دید کلی به گراف دارد نیز باعث بهبود عملکرد و کیفیت خوشه‌بندی می‌شود. دلیل مدت زمان بالای اجرای این الگوریتم نیز همین است که به دلیل محاسبات غیرخطی و پارامترهای نسبتاً زیاد و در نظر گرفتن زیر گراف‌های محلی و دید کلی به

گراف مدت زمان اجرا الگوریتم نیز افزایش می‌یابد.

از دیگر نکات حائز اهمیت تعیین تعداد خوشه‌هاست. با افزایش و یا کاهش خوشه‌ها، عملکرد و کیفیت خوشه‌بندی به شدت کاهش می‌یابد که این امر به نوعی نقطه ضعف استفاده از این روش خوشه‌بندی است که برای پوشش این ضعف می‌توان از روش‌های خوشه‌بندی سلسله مراتبی استفاده نمود.

۴-۷ تحلیل خوشه‌ها

در این مرحله به تحلیل هریک از خوشه‌ها در گراف‌ها می‌پردازیم و اطلاعات کلی در مورد خوشه‌ها پیدا می‌کنیم. برای تحلیل خوشه‌ها در هر مرحله با توجه به مقادیر میانگین تعداد معاملات و میانگین ارزش ریالی معاملات فرض می‌کنیم که اگر تأمین‌کننده‌ای در تعداد معاملات کمتر (کمتر از میانگین)، ارزش ریالی معاملاتی بالایی (بالاتر از حد میانگین ارزش ریالی معاملات) داشته باشد، احتمال مشکوک بودن آن تأمین‌کننده وجود دارد. عامل مهم دیگر در شناسایی و تحلیل خوشه‌ها مقدار معیار مرکزیت نزدیکی است که هرچه این مقدار بیشتر باشد به معنی ارتباط بیشتر با گره‌های دیگر و افزایش احتمال تبانی در مناقصه است.

جدول ۰-۷: تحلیل داده‌های خوشه‌های گراف بدون جهت با وزن تعداد معاملات و ۴۱۷ گره

شناسه‌ی خوشه	تعداد نمونه‌ها در هر خوشه	میانگین مقادیر در هر خوشه و اطلاعات آن‌ها	تعداد داده‌ها با رفتار مشکوک	تعداد داده‌های با رفتار معمولی
خوشه ۰	۳۵۴ نمونه(۸ دستگاه اجرایی و ۳۴۶ تأمین‌کننده)	میانگین تعداد معاملات در خوشه ۵ میانگین ارزش ریالی معاملات در این خوشه:۶۲ میلیارد ریال کمترین مقدار ارزش ریالی معاملات:۴۸۰ هزار ریال بیشترین مقدار ارزش ریالی معاملات:۳ هزار و ۵۰۰ میلیارد ریال	۲۳ نمونه	۳۳۱ نمونه
خوشه ۱	۱ نمونه (دستگاه اجرایی)	با توجه به فاصله این خوشه از دیگر خوشه‌ها، این خوشه که داده‌ی کمی نیز دارد، به عنوان داده نویز محسوب می‌شود و می‌تواند حذف بشود و یا با تغییر مقدار خوشه‌ها در دیگر خوشه‌ها قرار بگیرد.		
خوشه ۲	۱۲ نمونه(۴ تأمین‌کننده و ۸ دستگاه اجرایی)	میانگین تعداد معاملات در خوشه ۲۸ میانگین ارزش ریالی معاملات در این خوشه:۶۰۹ میلیارد ریال کمترین مقدار ارزش ریالی معاملات:۲۹ میلیارد ریال بیشترین مقدار ارزش ریالی معاملات:۱,۹۰۰ میلیارد ریال	۰نمونه	۱۲ نمونه
خوشه ۳	یک نمونه(دستگاه اجرایی)	با توجه به فاصله این خوشه از دیگر خوشه‌ها، این خوشه که داده‌ی کمی نیز دارد، به عنوان داده نویز محسوب می‌شود و می‌تواند حذف بشود و یا با تغییر مقدار خوشه‌ها در دیگر خوشه‌ها قرار بگیرد.		

خوشه ۴	۴۹نمونه(۱۷ دستگاه اجرایی و ۳۲ تأمین‌کننده)	میانگین تعداد معاملات در خوشه: ۱۱ میانگین ارزش ریالی معاملات در این خوشه:۲۲۶ میلیارد ریال کمترین مقدار ارزش ریالی معاملات:۴,۱۴۷ میلیارد ریال بیشترین مقدار ارزش ریالی معاملات:۳,۵۱۰ میلیارد ریال	۰نمونه	۴۹ نمونه

جدول ۰-۸: تحلیل داده‌های خوشه‌های گراف بدون جهت با وزن تعداد معاملات و ۱۱۱۲ گره

شناسه‌ی خوشه	تعداد نمونه‌ها در هر خوشه	میانگین مقادیر در هر خوشه و اطلاعات آن‌ها	تعداد داده‌ها با رفتار مشکوک	تعداد داده‌های با رفتار معمولی
خوشه ۰	۱۰۵۰ نمونه(۷۹ دستگاه اجرایی و ۹۷۱ تأمین‌کننده)	میانگین تعداد معاملات در خوشه ۹ میانگین ارزش ریالی معاملات در این خوشه:۱۵۵ میلیارد ریال کمترین مقدار ارزش ریالی معاملات:۷۵ هزار ریال بیشترین مقدار ارزش ریالی معاملات:۳۱ هزار میلیارد ریال	۴۶ نمونه	۱۰۰۴نمونه
خوشه ۱	۲نمونه(۱ دستگاه اجرایی و ۱ تأمین‌کننده)	با توجه به فاصله این خوشه از دیگر خوشه‌ها، این خوشه که داده‌ی کمی نیز دارد، به عنوان داده نویز محسوب می‌شود و می‌تواند حذف بشود و یا با تغییر مقدار خوشه‌ها در دیگر خوشه‌ها قرار بگیرد.		

خوشه۲	۵۲نمونه(۳۳ دستگاه اجرایی و ۱۹تأمین‌کننده)	میانگین تعداد معاملات در خوشه ۴۳ میانگین ارزش ریالی معاملات در این خوشه:۱۷۴۹میلیارد ریال کمترین مقدار ارزش ریالی معاملات:۲۷ میلیارد ریال بیشترین مقدار ارزش ریالی معاملات:۹۹۱۰ میلیارد ریال	۴نمونه	۴۸نمونه
خوشه۳	۸نمونه(۴دستگاه اجرایی و ۲ تأمین‌کننده)	میانگین تعداد معاملات در خوشه ۱۰۱ میانگین ارزش ریالی معاملات در این خوشه:۴۸۰۰ میلیارد ریال کمترین مقدار ارزش ریالی معاملات:۱۴۲ میلیارد ریال بیشترین مقدار ارزش ریالی معاملات:۱۵۱۰۰میلیارد ریال	۰ نمونه	۸ نمونه

جدول ۰-۹:تحلیل داده‌های خوشه‌های گراف بدون جهت با وزن ارزش ریالی معاملات و ۸۷۶ گره

شناسه‌ی خوشه	تعداد نمونه‌ها در هر خوشه	میانگین مقادیر در هر خوشه و اطلاعات آن‌ها	تعداد داده‌ها با رفتار مشکوک	تعداد داده‌های با رفتار معمولی
خوشه۰	۸۲۲نمونه(۶۸ دستگاه اجرایی و ۷۵۴ تأمین‌کننده)	میانگین تعداد معاملات در خوشه: ۹ میانگین ارزش ریالی معاملات در این خوشه:۷۷ میلیارد ریال کمترین مقدار ارزش ریالی معاملات:۱۲۴ هزار ریال بیشترین مقدار ارزش ریالی معاملات:۶۵۶ میلیارد	۲۷ نمونه	۷۹۵ نمونه

خوشه ۱	۱۶نمونه (۱۱نمونه دستگاه اجرایی و ۵نمونه تأمین‌کننده)	میانگین تعداد معاملات در خوشه ۷۵ میانگین ارزش ریالی معاملات در این خوشه:۴۹۰۰میلیارد ریال کمترین مقدار ارزش ریالی معاملات:۳۳۰۰ میلیارد ریال بیشترین مقدار ارزش ریالی معاملات:۶۹۰۰ میلیارد ریال	۲ نمونه	۱۴ نمونه
خوشه ۲	۴ نمونه (۳ دستگاه اجرایی و ۱ تأمین‌کننده)	با توجه به فاصله این خوشه از دیگر خوشه‌ها، این خوشه که داده‌ی کمی نیز دارد، به عنوان داده نویز محسوب می‌شود و می‌تواند حذف بشود و یا با تغییر مقدار خوشه‌ها در دیگر خوشه‌ها قرار بگیرد		
خوشه ۳	۳۴نمونه (۹ دستگاه اجرایی و ۲۵ تأمین‌کننده)	میانگین تعداد معاملات در خوشه ۳۷ میانگین ارزش ریالی معاملات در این خوشه:۱۳۰۰ میلیارد ریال کمترین مقدار ارزش ریالی معاملات:۷۰۹ میلیارد ریال بیشترین مقدار ارزش ریالی معاملات:۲۹۰۰ میلیارد ریال	۲ نمونه	۳۲ نمونه

جدول ۰-۱۰:تحلیل داده‌های خوشه‌های گراف بدون جهت با وزن ارزش ریالی معاملات و۸۳ گره

شناسه‌ی خوشه	تعداد نمونه‌ها در هر خوشه	میانگین مقادیر در هر خوشه و اطلاعات آن‌ها	تعداد داده‌ها با رفتار مشکوک	تعداد داده‌های با رفتار معمولی
خوشه۰	۷۲نمونه(۳ دستگاه اجرایی و ۶۹ تأمین‌کننده)	میانگین تعداد معاملات در خوشه ۱۰ میانگین ارزش ریالی معاملات در این خوشه:۴۶ هزار میلیارد کمترین مقدار ارزش ریالی معاملات:۸۳۰ هزار ریال بیشترین مقدار ارزش ریالی معاملات:۱۹۰ میلیارد ریال	۱۱ نمونه	۶۱نمونه

خوشه۱	۱نمونه(دستگاه اجرایی)	با توجه به فاصله این خوشه از دیگر خوشه‌ها، این خوشه که داده‌ی کمی نیز دارد، به عنوان داده نویز محسوب می‌شود و می‌تواند حذف بشود و یا با تغییر مقدار خوشه‌ها در دیگر خوشه‌ها قرار بگیرد.		
خوشه۲	۱۰نمونه (۴ دستگاه اجرایی و ۶ تأمین‌کننده)	میانگین تعداد معاملات در خوشه ۳۹ میانگین ارزش ریالی معاملات در این خوشه:۴۰۹ میلیارد ریال کمترین مقدار ارزش ریالی معاملات:۲۵۳ میلیارد ریال بیشترین مقدار ارزش ریالی معاملات:۹۲۷ میلیارد ریال	۱نمونه	۹نمونه

تحلیل جداول بالا با توجه به مقدار دقیق تعداد خوشه‌هاست و در درون هر خوشه و داده‌های آن‌ها به شناسایی رفتار ناهنجار و الگوی متفاوت داده‌ها پرداختیم، اما با توجه به اینکه در الگوریتم خوشه‌بندی $KMeans$ باید تعداد خوشه‌ها از قبل نیز مشخص باشد، با توجه به مقدار دقیق خوشه‌ها، دریافتیم که با افزایش تعداد خوشه‌ها، تعدادی از خوشه‌ها نیز از دیگر خوشه‌ها دورتر هستند و تعداد داده‌ی کمتری نیز دارند که این حاکی از رفتار و الگوی متفاوت آن خوشه‌هاست. در واقع تحلیل را به دو صورت میان‌خوشه‌ای و بین‌خوشه‌ای انجام داده‌ایم. در تحلیل بین‌خوشه‌ای، خوشه‌هایی که دارای رفتار متفاوتی هستند نیز با تعداد کمی داده مشخص می‌شوند و این بار برچسب مشکوک بودن را به خوشه‌ها نسبت می‌دهیم. در تمام موارد برای شناسایی خوشه‌هایی با رفتار مشکوک، تعداد خوشه‌ها را دو برابر مقداری که از الگوریتم Elbow Method به دست آوردیم، در نظر گرفتیم.

جدول ۰-۱۱: تحلیل خوشه‌های گراف بدون جهت با وزن تعداد معاملات و ۴۱۷ گره

تعداد خوشه‌ها	تعداد نمونه‌های خوشه	رفتار مشکوک	رفتار نرمال
خوشه۰	۹	●	
خوشه۱	۳۱۳		●
خوشه۲	۱	●	
خوشه۳	۱	●	

خوشه۴	۳	●	
خوشه۵	۸	●	
خوشه۶	۲۴		●
خوشه۷	۴۱		
خوشه۸	۱	●	
خوشه۹	۱۶		●

جدول ۰-۱۲:تحلیل خوشه‌های گراف بدون جهت با وزن تعداد معاملات و ۱۱۱۲ گره

تعداد خوشه‌ها	تعداد نمونه‌های خوشه	رفتار مشکوک	رفتار نرمال
خوشه۰	۹۵۴		●
خوشه۱	۲	●	
خوشه۲	۲۱		●
خوشه۳	۱	●	
خوشه۴	۳۱		●
خوشه۵	۱	●	
خوشه۶	۶	●	
خوشه۷	۹۶		●

جدول ۰-۱۳:تحلیل خوشه‌های گراف بدون جهت با وزن ارزش ریالی معاملات و ۸۷۶ گره

تعداد خوشه‌ها	تعداد نمونه‌های خوشه	رفتار مشکوک	رفتار نرمال
خوشه۰	۷۴۳		●
خوشه۱	۱۲		●
خوشه۲	۳	●	
خوشه۳	۲۴		●
خوشه۴	۱	●	
خوشه۵	۶	●	
خوشه۶	۲	●	
خوشه۷	۸۵		●

جدول ۰-۱۴:تحلیل خوشه‌های گراف بدون جهت با وزن ارزش ریالی معاملات و ۸۳ گره

تعداد خوشه‌ها	تعداد نمونه‌های خوشه	رفتار مشکوک	رفتار نرمال
خوشه۰	۴	●	
خوشه۱	۵۲		●
خوشه۲	۱	●	
خوشه۳	۱	●	
خوشه۴	۲۰		●
خوشه۵	۵	●	

همانطور که در فرضیات فصل یک بحث شد، خوشه‌هایی که از دیگر خوشه‌ها دورتر هستند به عنوان داده ناهنجار تعریف می‌شوند. فرضیه بعدی دیگر، درون هر خوشه بود که هرچه درون یک خوشه داده‌ها از مرکز خوشه دورتر باشند به عنوان داده با رفتار مشکوک در نظر گرفته می‌شوند. همانطور که در جداول ۴-۱۱ تا ۴-۱۴ مشاهده کردیم، با افزایش تعداد خوشه‌ها، داده‌هایی در خوشه‌هایی دورتر قرار می‌گیرند که وجود این خوشه‌ها حاکی از رفتار و الگوی متفاوت داده‌های آن‌هاست.

۴-۸ جمع‌بندی

در این فصل با توجه به الگوریتم ارائه شده در فصل سوم به بررسی نتایج و انجام پیاده‌سازی‌های مربوطه پرداخته‌ایم. ابتدا با استفاده از تکنیک‌های گراف‌کاوی به استخراج اطلاعات مفید از گراف پرداختیم و انجمن‌های باارزش را از گراف پیدا کردیم، سپس برای اجرای الگوریتم‌های یادگیری‌های ماشین و داده‌کاوی که همان خوشه‌بندی است، گراف‌هایی که به دست آمده‌اند و به دو گراف با وزن ارزش ریالی معاملات و تعداد معاملات بودند را به عنوان ورودی به الگوریتم‌های تعبیه گراف دادیم تا بازنمایی از گره‌های گراف در ابعاد با ویژگی‌های بیشتر دستیابی داشته باشیم. در مرحله‌ی آخر با استفاده از اجرای الگوریتم خوشه‌بندی بر روی داده‌های تعبیه‌شده گراف و ویژگی‌های اصلی گراف که همان تعداد معاملات و ارزش ریالی معاملات است به خوشه‌بندی پرداختیم. درنهایت با استفاده از معیارهای ارزیابی برای هر خوشه و تحلیل خوشه‌ها و داده‌های هر خوشه به بررسی و شناسایی داده‌های با رفتار متفاوت و

الگوی متفاوت پرداختیم.

فصل ۵ نتیجه گیری و پژوهش های آتی

۵-۱ نتیجه‌گیری و یافته‌های پژوهش

در این پژوهش، با استفاده از ترکیب روش‌های گراف‌کاوی، یادگیری عمیق و تکنیک یادگیری ماشین به شناسایی رفتار مشکوک داده‌ها پرداختیم. همانطور که در فصل اول به این نکته اشاره شد، استفاده از روش‌های سنتی و پیشین در مورد داده‌های تراکنش‌های مالی و معاملات و خریدوفروش‌های بخش دولتی چالش‌های زیادی را به همراه داشت و به دلیل شفافیت پایین، دولت‌ها به سمت استفاده از فناوری اطلاعات و ارتباطات در راستای انجام الکترونیکی خریدوفروش‌ها و انجام مناقصات قدم برداشتند. بعد از استقرار دولت الکترونیک و استفاده کشورها از این مفهوم و انجام معاملات در بستر وب، با توجه به افزایش شفافیت، تسهیل امور معاملات و دیگر فواید آن، احتمال فساد و تبانی نیز به طبع آن افزایش یافته است. در واقع از آنجایی که انجام معاملات شکل دیجیتالی به خود گرفته، فساد نیز در این بستر دور از ذهن نبوده است.

در فرآیند معاملات دولتی به دلیل هزینه‌های زیاد انجام شده، حجم بالای تراکنش‌ها، پیچیدگی فرآیند انجام معاملات و تعداد سرمایه‌گذاران؛ فضا و بستر برای ایجاد تبانی و فساد اقتصادی مهیا است. به دلیل ماهیت رابطه‌ای بودن داده‌ها در دنیای واقعی تراکنش‌ها و معاملات خریدوفروش، در این پژوهش در وهله‌ی اول با استفاده از تکنیک‌های گراف‌کاوی بعد از مدل‌سازی گراف به شناسایی و استخراج اطلاعات مفید در مورد گراف معاملات بین مناقصه‌گران و تأمین‌کنندگان پرداختیم. گراف معاملات یکبار با وزن تعداد معاملات و یکبار با وزن یال ارزش ریالی معاملات در نظر گرفتیم. در این مرحله به دلیل حجم بالای گراف و سربار محاسباتی موجود، به شناسایی و استخراج انجمن‌های باارزش در گراف پرداختیم که حاصل این کار دو انجمن در بخش گراف با وزن ارزش ریالی معاملات، و دو انجمن در بخش گراف با وزن تعداد معاملات انجام شده بود.

بعد از استخراج انجمن‌های پرارزش، در این مرحله به دلیل عملکرد بسیار مناسب و رویکردهای جدید یادگیری عمیق، به تعبیه گره‌های گراف‌های انجمن‌ها پرداختیم. بدین معنی که برای هر گره از گراف

بردار ویژگی‌ای را به دست آورده‌ایم که اطلاعات دیگر گره‌ها را نیز به همراه داشت. پس از اجرای سه الگوریتم تعبیه گراف، بردارهای ویژگی به دست آمده را به همراه ویژگی‌های اصلی گراف و اطلاعات و معیارها و شاخص‌هایی که از گراف به دست آوردیم، به عنوان ورودی به الگوریتم‌های خوشه‌بندی K-Means دادیم. در مرحله‌ی مقایسه و ارزیابی برای کیفیت هر خوشه از معیار Silhoutte استفاده کردیم که در این مسئله به دلیل اندازه خوشه‌های متفاوت معیار مطلوب‌تری بوده است.

برای بررسی عملکرد روش پیشنهادی، از داده‌های دنیای واقعی و معاملات بین مناقصه‌گران و تأمین‌کنندگان که در سطح کشور ایران و در بازه زمانی یازده سال بوده استفاده شده است. تعداد ۱۲۸۰۸ تأمین‌کننده و ۱۳۹۸ مناقصه‌گزار مورد بررسی قرار گرفته‌اند. در مرحله‌ی خوشه‌بندی دریافته‌ایم که استفاده از روش‌های تعبیه گراف به دلیل اینکه از دیگر اطلاعات گره‌های گراف نیز در هر گره برای خوشه‌بندی استفاده می‌کند عملکرد بهتری داشته است و میانگین معیار کیفیت خوشه‌بندی (Silhoutte) قبل و بعد از استفاده از این بردار ویژگی در کنار دیگر داده‌ها تغییراتی مثبتی داشته است. یکی از یافته‌های مهم پژوهش این است که هرچه ابعاد تعبیه گراف و طول بردار ویژگی طولانی‌تر باشد، در واقع اطلاعات بیشتری را در مورد گره‌ها و همسایگان آن‌ها در اختیار می‌گذارد. علاوه بر این نیز پیچیدگی زمانی و فضایی بیشتری را دارد. در واقع با تغییر پارامتر ابعاد بردار ویژگی در الگوریتم‌های تعبیه گراف و دریافتیم که هرچه اطلاعات بیشتری در مورد گره‌های دیگر و همسایگان آن‌های در اختیار الگوریتم‌های خوشه‌بندی باشد، خوشه‌های باکیفیت بهتری از هم جدا می‌شوند.

روش ارائه‌شده در گراف‌ها حاکی از عملکرد مناسب استفاده از روش‌های یادگیری عمیق و روش‌های تعبیه گراف بوده است. همانطور که در آزمایش‌های نشان داده شده است، تحلیل خوشه‌بندی و اجرای خوشه‌بندی وابستگی زیادی به تعبیه گراف دارد.

در تحلیل خوشه‌ها با استفاده از فرضیات موجود به بررسی رفتار و الگوی مشکوک داده‌ها پرداختیم. استفاده از ویژگی‌ها و مقادیر و اطلاعات هر خوشه اطلاعات مناسبی را برای تصمیم‌گیری در مورد الگوی

داده‌ها ناهنجار به ما داده است.

۵-۲ پیشنهادات برای کارهای آتی

با توجه به اینکه نتایج در مورد این مسئله وابستگی‌های زیادی به پارامترهای مختلف دارد و استفاده از روش‌های تعبیه گراف همچنان مورد بحث و بررسی محققان زیادی است، انجام مطالعات و تحقیقات بیشتر در این زمینه امری ضروری است. برای انجام مطالعات بیشتر می‌توان به پیشنهادهای زیر توجه کرد:

- استفاده از ویژگی‌های دیگر گراف معاملات در صورت در دسترس بودن آن‌ها. به عنوان مثال استفاده از ویژگی‌هایی مانند تاریخ انجام خریدوفروش، تاریخ انتشار سند مناقصه، تاریخ برگزاری مناقصات و تعداد برندگان هر مناقصه خریدوفروش. با دانستن ویژگی‌های دیگر خوشه‌بندی بهتری را می‌توان بر روی داده‌ها انجام داد و رفتار مشکوک افراد در معاملات با دقت بالاتری قابل شناسایی است.
- برچسب‌گذاری و در دسترس بودن داده‌هایی که رفتار مشکوک و یا معمولی آن‌ها مشخص باشد باعث عملکرد بهتر در طبقه‌بندی و شناسایی دقیق‌تر داده‌های جدید می‌شود. همانطور که گفته شد، به علت یادگیری بدون نظارت داده‌ها
- انجام تعبیه گراف بر روی یال‌ها. در پژوهش حاضر روش‌های یادگیری بازنمایی گراف بر روی گره‌های انجام شده که ویژگی‌های هر گره و گره همسایه را در نظر می‌گرفت. برای انجام تعبیه گراف بر روی یال‌ها باید ویژگی‌های را بر روی یال‌ها تعریف شود و از گراف دانش استفاده شود که باعث عملکرد بهتر و یادگیری بهتر گراف می‌شود.
- انجام تعبیه گراف بر روی زیرگراف‌های هر گراف و رویکردهایی مانند Pathcy-san و Sub2vec. با انجام تعبیه گراف بر روی زیرگراف‌ها علاوه بر دقت بهتر می‌توان مدت زمان اجرای الگوریتم را با استفاده از شناسایی انجمن ها و زیرگراف‌های باارزش‌تر کاهش داد. در

واقع شناسایی و یادگیری انجمن‌ها باارزش نیز به صورت موازی باهم انجام می‌شود.

- شناسایی انجمن‌های همپوشان و جداسازی و استخراج هر یک انجمن‌ها
- انجام پیش‌بینی ارتباط بین گره‌ها (link prediction) برای شناسایی ارتباط بین مناقصه‌گزاران و مناقصه گران.
- تغییر پارامترها و بهینه‌سازی هرکدام از الگوریتم‌های تعبیه گراف. به عنوان مثال در الگوریتم DeepWalk، مقدار بردار ویژگی با تغییر تعداد گام‌های تصادفی و طول گام و اندازه پنجره Skip-gram که باعث عملکرد بهتر در یادگیری گره‌های همسایه می‌شود.
- استفاده از روش‌های دیگر برای تعبیه گراف مانند فاکتورسازی ماتریس و روش‌های شبکه کانولوشنی گراف و مقایسه روش‌ها با یکدیگر. برخی از
- استفاده از الگوریتم‌های دیگر خوشه‌بندی مانند خوشه‌بندی مبتنی بر تراکنش و الگوریتم خوشه‌بندی سلسله مراتبی که در مدت زمان کمتر و با پارامتر کمتر و دقت بیشتر می‌تواند داده‌ها در خوشه‌های مختلف دسته بندی کند.
- اختصاص مقدار و امتیاز به هر داده خوشه و شناسایی داده‌های ناهنجار براساس امتیاز داده‌ها.

مراجع

[1] N. R. Prasad, S. Almanza-Garcia, and T. T. Lu, “Anomaly detection,” *Comput. Mater. Contin.*, vol. 14, no. 1, pp. 1–22, Jul. 2009, doi: 10.1145/1541880.1541882.

[2] J. Lin, E. Keogh, A. Fu, and H. Van Herle, “Approximations to magic: finding unusual medical time series,” in *18th IEEE Symposium on Computer-Based Medical Systems (CBMS’05)*, 2005, pp. 329–334, doi: 10.1109/CBMS.2005.34.

[3] R. J. Bolton, R. J. Bolton, D. J. Hand, and D. J. H, “Unsupervised Profiling Methods for Fraud Detection,” *PROC. Credit SCORING Credit Control VII*, pp. 5--7, 2001, Accessed: Dec. 01, 2021. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.24.5743.

[4] D. J. Marchette, *Computer Intrusion Detection and Network Monitoring*. Springer New York, 2001.

[5] محمدیان، محمد و بهاری، آرمان،۱۳۹۵،اصلاح نظام اداری و جلوگیری از فساد در پرتو استقرار دولت الکترونیک،ششمین کنفرانس بین المللی حسابداری و مدیریت با رویکرد علوم پژوهشی نوین،تهران.

[6] رستمی چراتی، محمدرضا و رضازاده قاجاری، مقدم و سلیمانی، سیده فاطمه،۱۳۹۴،سیستم یکپارچه معاملات دولتی،چهارمین کنفرانس ملی مدیریت و حسابداری،تهران

[7] هنرآموز، سحر و سلاجقه، سنجر،۱۳۹۱،دولت الکترونیک از تئوری تا عمل، راهبرد توسعه، شماره ۲۹

[8] شاکری، اقبال و ابراهیم نژاد، میثم و رستمی، حسن،۱۳۹۲، بررسی میزان هماهنگی قانون برگزاری مناقصات با موافقتنامه معاملات دولتی سازمان تجارت جهانی، دومین کنفرانس ملی حسابداری, مدیریت مالی و سرمایه گذاری ،گرگان

[9] طالع نسب، وحید،۱۳۹۶، پیشگیری از فساد مالی در مناقصه ها و مزایده ها،سومین کنفرانس سراسری حقوق و مطالعات قضایی

[10] V. Van Vlasselaer *et al.*, “APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions,” *Decis. Support Syst.*, vol. 75, pp. 38–48, Jul. 2015, doi: 10.1016/j.dss.2015.04.013.

[11] D. J. Hand, C. Whitrow, N. M. Adams, P. Juszczak, and D. Weston, “Performance criteria for plastic card fraud detection tools,” *J. Oper. Res. Soc.*, vol. 59, no. 7, pp. 956–962, 2008, doi: 10.1057/palgrave.jors.2602418.

[12] N. Carneiro, G. Figueira, and M. Costa, “A data mining based system for credit-card fraud detection in e-tail,” *Decis. Support Syst.*, vol. 95, pp. 91–101, Mar. 2017, doi: 10.1016/j.dss.2017.01.002.

[13] طجرلو، رضا.، قربانی درآباد، بهزاد.، ۱۳۹۵. سلامتی در معاملات دولتی. مطالعات حقوق انرژی

[14] W. Xu and Y. Liu, “An optimized SVM model for detection of fraudulent online credit card transactions,” *Proc. - 2012 Int. Conf. Manag. e-Commerce e-Government, ICMeCG 2012*, pp. 14–17, 2012, doi: 10.1109/ICMECG.2012.39.

[15] M. Zareapoor and P. Shamsolmoali, “Application of credit card fraud detection: Based on bagging ensemble classifier,” in *Procedia Computer Science*, Jan. 2015, vol. 48, no. C, pp. 679–685, doi: 10.1016/j.procs.2015.04.201.

[16] N. K. Gyamfi and J. D. Abdulai, “Bank Fraud Detection Using Support Vector Machine,” *2018 IEEE 9th Annu. Inf. Technol. Electron. Mob. Commun. Conf. IEMCON 2018*, pp. 37–41, Jan. 2019, doi: 10.1109/IEMCON.2018.8614994.

[17] D. Qingshan, “Application of support vector machine in the detection of fraudulent financial statements,” *Proc. 2009 4th Int. Conf. Comput. Sci. Educ. ICCSE 2009*, pp. 1056–1059, 2009, doi: 10.1109/ICCSE.2009.5228542.

[18] T. K. Behera and S. Panigrahi, “Credit Card Fraud Detection: A Hybrid Approach Using Fuzzy Clustering & Neural Network,” in *Proceedings - 2015 2nd IEEE International Conference on Advances in Computing and Communication Engineering, ICACCE 2015*, Oct. 2015, pp. 494–499, doi: 10.1109/ICACCE.2015.33.

[19] M. R. Haratinik, M. Akrami, S. Khadivi, and M. Shajari, “FUZZGY: A hybrid model for credit card fraud detection,” *2012 6th Int. Symp. Telecommun. IST 2012*, pp. 1088–1093, 2012, doi: 10.1109/ISTEL.2012.6483148.

[20] A. Agrawal, S. Kumar, and A. K. Mishra, “Credit Card Fraud Detection: A case study,” in *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, 2015, pp. 5–7.

[21] A. Khan, T. Singh, and A. Sinhal, “Implement credit card fraudulent detection system using observation probabilistic in hidden Markov model,” *undefined*, 2012, doi: 10.1109/NUICONE.2012.6493206.

[22] D. Iyer, A. Mohanpurkar, S. Janardhan, D. Rathod, and A. Sardeshmukh, “Credit card fraud detection using hidden Markov model,” in *Proceedings of the 2011 World Congress on Information and Communication Technologies, WICT 2011*, 2011, pp. 1062–1066, doi: 10.1109/WICT.2011.6141395.

[23] N. Malini and M. Pushpa, “Analysis on credit card fraud identification techniques based on KNN and outlier detection,” *Proc. 3rd IEEE Int. Conf. Adv. Electr. Electron. Information, Commun. Bio-Informatics, AEEICB 2017*, pp. 255–258, Jul. 2017, doi: 10.1109/AEEICB.2017.7972424.

[24] J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, “Credit card fraud detection using machine learning techniques: A comparative analysis,” *Proc. IEEE Int. Conf. Comput. Netw. Informatics, ICCNI 2017*, vol. 2017-January, pp. 1–9, Nov. 2017, doi: 10.1109/ICCNI.2017.8123782.

[25] P. Hajek and R. Henriques, “Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods,” *Knowledge-Based Syst.*, vol. 128, pp. 139–152, Jul. 2017, doi:

10.1016/J.KNOSYS.2017.05.001.

[26] A. Srivastava, M. Yadav, S. Basu, S. Salunkhe, and M. Shabad, “Credit card fraud detection at merchant side using neural networks,” in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 2016, pp. 667–670.

[27] F. Ghobadi and M. Rohani, “Cost sensitive modeling of credit card fraud using neural network strategy,” *Proc. - 2016 2nd Int. Conf. Signal Process. Intell. Syst. ICSPIS 2016*, Mar. 2017, doi: 10.1109/ICSPIS.2016.7869880.

[28] Y. Sahin and E. Duman, “Detecting credit card fraud by ANN and logistic regression,” *INISTA 2011 - 2011 Int. Symp. Innov. Intell. Syst. Appl.*, pp. 315–319, 2011, doi: 10.1109/INISTA.2011.5946108.

[29] F. Anowar and S. Sadaoui, “Detection of Auction Fraud in Commercial Sites,” vol. 15, no. 1, 2020, doi: 10.4067/S0718-18762020000100107.

[30] S. J. Ester M, Kriegel H P, “A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of Second International Conference on Knowledge Discovery and Data Mining,” *Kdd*, vol. 96, no. 34, pp. 226–231, 1996, Accessed: Dec. 11, 2021. [Online]. Available: https://dl.acm.org/doi/10.5555/3001460.3001507.

[31] M. Ahmed and A. N. Mahmood, “A novel approach for outlier detection and clustering improvement,” in *Proceedings of the 2013 IEEE 8th Conference on Industrial Electronics and Applications, ICIEA 2013*, 2013, pp. 577–582, doi: 10.1109/ICIEA.2013.6566435.

[32] V. Hautamäki, S. Cherednichenko, I. Kärkkäinen, T. Kinnunen, and P. Fränti, “Improving K-means by outlier removal,” in *Lecture Notes in Computer Science*, 2005, vol. 3540, pp. 978–987, doi: 10.1007/11499145_99.

[33] Z. He, X. Xu, and S. Deng, “Discovering cluster-based local outliers,” *Pattern Recognit. Lett.*, vol. 24, no. 9–10, pp. 1641–1650, Jun. 2003, doi: 10.1016/S0167-8655(03)00003-5.

[34] H. Issa and M. A. Vasarhelyi, “Application of Anomaly Detection Techniques to Identify Fraudulent Refunds,” *SSRN Electron. J.*, Aug. 2011, doi: 10.2139/SSRN.1910468.

[35] S. Thiprungsri and M. A. Vasarhelyi, “Cluster analysis for anomaly detection in accounting data: An audit approach,” *Int. J. Digit. Account. Res.*, vol. 11, pp. 69–84, 2011, doi: 10.4192/1577-8517-v11_4.

[36] W. H. Chang and J. S. Chang, “Using clustering techniques to analyze fraudulent behavior changes in online auctions,” *ICNIT 2010 - 2010 Int. Conf. Netw. Inf. Technol.*, pp. 34–38, 2010, doi: 10.1109/ICNIT.2010.5508564.

[37] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software,” *ACM SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009, doi: 10.1145/1656274.1656278.

[38] N. A. Le Khac and M. T. Kechadi, “Application of data mining for anti-money laundering detection: A case study,” *Proc. - IEEE Int. Conf. Data Mining, ICDM*,

pp. 577–584, 2010, doi: 10.1109/ICDMW.2010.66.

[39] M. Jans, N. Lybaert, and K. Vanhoof, “Data Mining for Fraud Detection : Toward an Improvement on Internal Control Systems ?,” *Int. Res. Symp. Account. Inf. Syst.*, pp. 1–17, 2006.

[40] S. Panigrahi, A. Kundu, S. Sural, and A. K. Majumdar, “Credit card fraud detection: A fusion approach using Dempster-Shafer theory and Bayesian learning,” *Inf. Fusion*, vol. 10, no. 4, pp. 354–363, Oct. 2009, doi: 10.1016/j.inffus.2008.04.001.

[41] L. Torgo and E. Lopes, “Utility-based fraud detection,” in *IJCAI International Joint Conference on Artificial Intelligence*, 2011, pp. 1517–1522, doi: 10.5591/978-1-57735-516-8/IJCAI11-255.

[42] R. Wang *et al.*, “Deep Learning for Anomaly Detection,” *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, no. February, pp. 3569–3570, 2020, doi: 10.1145/3394486.3406481.

[43] S. L. Domingos, R. N. Carvalho, R. S. Carvalho, and G. N. Ramos, “Identifying it purchases anomalies in the Brazilian Government Procurement System using deep learning,” *Proc. - 2016 15th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2016*, no. Cic, pp. 722–727, 2017, doi: 10.1109/ICMLA.2016.106.

[44] T. Amarbayasgalan, B. Jargalsaikhan, and K. H. Ryu, “Unsupervised novelty detection using deep autoencoders with density based clustering,” *Appl. Sci.*, vol. 8, no. 9, 2018, doi: 10.3390/app8091468.

[45] M. Zamini and G. Montazer, “Credit Card Fraud Detection using autoencoder based clustering,” *9th Int. Symp. Telecommun. With Emphas. Inf. Commun. Technol. IST 2018*, pp. 486–491, 2019, doi: 10.1109/ISTEL.2018.8661129.

[46] W. Min, W. Liang, H. Yin, Z. Wang, M. Li, and A. Lal, “Explainable Deep Behavioral Sequence Clustering for Transaction Fraud Detection,” 2021, [Online]. Available: http://arxiv.org/abs/2101.04285.

[47] S. Misra, S. Thakur, M. Ghosh, and S. K. Saha, “An Autoencoder Based Model for Detecting Fraudulent Credit Card Transaction,” in *Procedia Computer Science*, Jan. 2020, vol. 167, pp. 254–262, doi: 10.1016/j.procs.2020.03.219.

[48] N. T. N. Anh, T. Q. Khanh, N. Q. Dat, E. Amouroux, and V. K. Solanki, “Fraud detection via deep neural variational autoencoder oblique random forest,” Sep. 2020, doi: 10.1109/HYDCON48903.2020.9242753.

[49] J. H. Oh, J. Y. Hong, and J. G. Baek, “Oversampling method using outlier detectable generative adversarial network,” *Expert Syst. Appl.*, vol. 133, pp. 1–8, Nov. 2019, doi: 10.1016/j.eswa.2019.05.006.

[50] Y. J. Zheng, X. H. Zhou, W. G. Sheng, Y. Xue, and S. Y. Chen, “Generative adversarial network based telecom fraud detection at the receiving bank,” *Neural Networks*, vol. 102, pp. 78–86, Jun. 2018, doi: 10.1016/j.neunet.2018.02.015.

[51] I. Benchaji, S. Douzi, B. El Ouahidi, and J. Jaafari, “Enhanced credit card fraud detection based on attention mechanism and LSTM deep model,” *J. Big Data*, vol. 8, no. 1, p. 151, 2021, doi: 10.1186/s40537-021-00541-8.

[52] D. Savage, X. Zhang, X. Yu, P. Chou, and Q. Wang, “Anomaly detection in online

social networks," *Soc. Networks*, vol. 39, no. 1, pp. 62–70, Oct. 2014, doi: 10.1016/j.socnet.2014.05.002.

[53] R. B. Velasco, I. Carpanese, R. Interian, O. C. G. Paulo Neto, and C. C. Ribeiro, "A decision support system for fraud detection in public procurement," *Int. Trans. Oper. Res.*, vol. 28, no. 1, pp. 27–47, 2021, doi: 10.1111/itor.12811.

[54] J. Zhu, B. Wang, L. Li, and J. Wang, "Bidder Network Community Division and Collusion Suspicion Analysis in Chinese Construction Projects," *Adv. Civ. Eng.*, vol. 2020, 2020, doi: 10.1155/2020/6612848.

[55] G. C. G. van Erven, M. Holanda, and R. N. Carvalho, "Detecting evidence of fraud in the Brazilian government using graph databases," *Adv. Intell. Syst. Comput.*, vol. 570, pp. 464–473, 2017, doi: 10.1007/978-3-319-56538-5_47.

[56] S. Sarswat, K. M. Abraham, and S. K. Ghosh, "Identifying collusion groups using spectral clustering," 2015, [Online]. Available: http://arxiv.org/abs/1509.06457.

[57] M. Dadfarnia, F. Adibnia, M. Abadi, and A. Dorri, "Incremental collusive fraud detection in large-scale online auction networks," *J. Supercomput.*, vol. 76, no. 9, pp. 7416–7437, 2020, doi: 10.1007/s11227-020-03170-9.

[58] A. Khazane *et al.*, "DeepTrax: Embedding graphs of financial transactions," *Proc. - 18th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2019*, pp. 126–133, 2019, doi: 10.1109/ICMLA.2019.00028.

[59] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," *33rd Int. Conf. Mach. Learn. ICML 2016*, vol. 1, pp. 740–749, 2016.

[60] Y. Pei, F. Lyu, W. van Ipenburg, and M. Pechenizkiy, "Subgraph anomaly detection in financial transaction networks," pp. 1–8, 2020, doi: 10.1145/3383455.3422548.

[61] D. S. H. Tam, W. C. Lau, B. Hu, Q. F. Ying, D. M. Chiu, and H. Liu, "Identifying Illicit Accounts in Large Scale E-payment Networks -- A Graph Representation Learning Approach," 2019, [Online]. Available: http://arxiv.org/abs/1906.05546.

[62] Z. Liu, Y. Dou, P. S. Yu, Y. Deng, and H. Peng, "Alleviating the Inconsistency Problem of Applying Graph Neural Network to Fraud Detection," in *SIGIR 2020 - Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, vol. 4, no. 20, pp. 1569–1572, doi: 10.1145/3397271.3401253.

[63] H. Wang *et al.*, "Live-Streaming Fraud Detection: A Heterogeneous Graph Neural Network Approach," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2021, pp. 3670–3678, doi: 10.1145/3447548.3467065.

[64] P. Cui, X. Wang, J. Pei, and W. Zhu, "A Survey on Network Embedding," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 5, pp. 833–852, 2019, doi: 10.1109/TKDE.2018.2849727.

[65] H. Cai, V. W. Zheng, and K. C. C. Chang, "A Comprehensive Survey of Graph Embedding: Problems, Techniques, and Applications," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 9, pp. 1616–1637, 2018, doi: 10.1109/TKDE.2018.2807452.

[66] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: Online learning of social

representations," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 701–710, doi: 10.1145/2623330.2623732.

[67] A. Grover and J. Leskovec, "Node2vec: Scalable feature learning for networks," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, vol. 13-17-Augu, pp. 855–864, doi: 10.1145/2939672.2939754.

[68] D. Wang, P. Cui, and W. Zhu, "Structural deep network embedding," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, vol. 13-17-Augu, pp. 1225–1234, doi: 10.1145/2939672.2939753.

[69] W. L. Hamilton, R. Ying, and J. Leskovec, "Representation Learning on Graphs: Methods and Applications," pp. 1–24, 2017, [Online]. Available: http://arxiv.org/abs/1709.05584.

[70] Y. Liu *et al.*, "Long- and short-term preference model based on graph embedding for sequential recommendation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Sep. 2020, vol. 12115 LNCS, pp. 241–257, doi: 10.1007/978-3-030-59413-8_20.

[71] Z. Zhang, P. Cui, and W. Zhu, "Deep Learning on Graphs: A Survey," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 8, pp. 1–1, 2020, doi: 10.1109/tkde.2020.2981333.

واژه‌نامه فارسی به انگلیسی

داده پرت	Outlier Data
شناسایی نو و جدید بودن داده	Novelty Detection
اختلالات در داده‌ها	Abnormality in data
داده‌های عجیب و نامانوس	Unexpected Data
داده ترتیبی	Sequential Data
ناهنجاری نقطه‌ای	Point Anomaly
ناهنجاری مفهومی	Contextual Anomaly
ناهنجاری تجمعی	Collective Anomaly
یادگیری بدون نظارت	Unsupervised Learning
یادگیری با نظارت	Supervised Learning
یادگیری نیمه نظارتی	Semi-Supervised Learning
نفوذ در شبکه	Network Intrusion
کلاهبرداری و تقلب در تراکنش‌های مالی	Financial Fraud
پولشویی	Money Laundering
تقلب در کارت اعتباری	Credit Card Fraud
تقلب در صورت حساب‌های مالی	Financial Statement Fraud
تقلب در مسائل مناقصه و مزایده	Procurement Fraud
دولت الکترونیک	E-Government

شاخص ادراک فساد	Corruption Perceptions Index (CPI)
تبانی	Collusion
ماشین بردار پشتیبان	Support Vector Machine
اَبَرصفحه	Hyperplane
منطق فازی	Fuzzy Logic
مدل مخفی مارکوف	Hidden Markov Model
نزدیک‌ترین همسایه	Nearest Neighbor
شبکه بیزین	Baysian Network
شبکه‌های عصبی مصنوعی	Artificial Neural Network
شبکه‌های خودسازمانده	Self Organizing Map
الگوریتم‌های خوشه‌بندی	Clustering Alogrithm
معیار اطلاعات بیزین	Baysian Information Criteria
الگوریتم خوشه‌بندی مبتنی بر چگالی	Density-based Spatial Clustering of Applications with Noise(DBSCAN)
رمزگذار-رمزگشا	Encoder-decoder
شبکه‌های خودرمزگذار	Autoencoder Network
یادگیری بازنمایی	Representation Learning
میانگین مربعات خطا	Mean Square Error
جنگل تصادفی احتمالاتی	Statistical random Forest
یادگیری تجمعی	Ensemble Learning
شبکه‌های متخاصم مولد	Generative Adversarial Network
مدل‌های رشته‌به‌رشته	Sequence-to-Sequence

شبکه‌های حافظه کوتاه مدت بلند	Long Short-Term Memory Network
تکنیک نمونه‌برداری بیش از حد اقلیت مصنوعی	Synthetic Minority Over-sampling TEchnique(SMOTE)
تعبیه‌گراف	Graph Embedding
یکریختی	Homophily
چندگانگی	Multiplexity
پیمانه‌ای بودن	Modularity
مرکزیت درجه	Degree Centerality
مرکزیت نزدیکی	Closeness Centerality
مرکزیت بینابینی	Betweenness Centerality
جستجوی سطحی اولین	Breadth First Search
جستجوی عمقی اولین	Depth First Search
بیش‌برازش	Overfitting

Abstract

Identifying the different behavior and pattern of data among the large amount of information available is of great importance. Nowadays, one of the most important areas for identifying anomalies in data is the financial field and conducting transactions and sales. Due to the high volume of transactions, the large number of investors and the complexity of the transaction process, fraud and economic collusion are among the problems that exist in this sector. With the establishment and growth of e-government in many countries, many tenders and auctions conducted in the public sector go beyond the traditional methods and are conducted in the context of information technology and electronically. The present study was conducted with the aim of clustering and identifying suspicious individuals in public sector tenders and has investigated the effect of using deep learning methods representing the trading graph in identifying suspicious data behavior. In this study, with the help of clustering method, customer behavior in transactions was analyzed and data pattern in different clusters was analyzed.

The set of data used in the present study includes tenders conducted by government agencies in Iran, which is available to the public in the period of 1389 to 1398 in the government procurement system. After modeling the data in the form of trading graphs, important graph indicators were extracted by network analysis methods and valuable trading graph associations were extracted and suppliers and executive agencies influencing large transactions were identified. Graph nodes were embedded using graph representation deep learning approaches to implement clustering algorithm and cluster analysis on associations data. The characteristics of each trading node (amount of value of the transaction and the number of transactions) along with the characteristics obtained through graph analysis and graph embedding feature vector, were considered as data input of the clustering algorithm. Finally, by analyzing the clusters and data within each of those clusters, suspicious data with different behavior were identified.

The research findings show that the use of graph properties along with the embedded feature vector of graph nodes obtained by deep learning methods of graph representation, contributed significantly to better and higher quality clustering of data. In addition, this study showed that the longer the feature vector in graph embedding, the better the data clustering performance due to the availability of more information of neighboring nodes in the graph, and the more accurate the data is identified with suspicious behavior among the data. .

Keywords: anomaly detection, e-government, graph analysis, graph embedding, graph representation learning, clustering, KMeans algorithm, deep learning

Shahrood University of Technology

Faculty of Computer Engineering and Information Technology

M.Sc Thesis in Artificial Intelligence Engineering

Fraud Detection in financial transactions using deep learning methods

By:Seyyed Ali Akbar Hosseini

Supervisor:
Dr. Hoda Mashayekhi

February 2022