فصلنامه علمی پژوهشی
دانش حسابداری و حسابرسی مدیریت
دوره ۱۷/ شماره ۱ (پیاپی ۶۵)/ بهار ۱۴۰۷
صفحه ۳۲۷ تا ۳۴۴

ردیابی استقرایی فرار مالیاتی: یک چارچوب ترکیبی مبتنی‌بر کشف جامعه و GraphSAGE

فاطمه عیوقی تفتی
دانشجوی دکتری، مدیریت فناوری اطلاعات گرایش کسب و کار هوشمند، دانشکده مدیریت و اقتصاد، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران.
f.ayoughitafti@iau.ac.ir
احمد ابراهیمی
دانشیار، عضو هیات علمی گروه مهندسی صنایع، دانشکده فنی و مهندسی، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران. (نویسنده مسئول)
ahmadebrahimi@iau.ac.ir
بابک فرهنگ مقدم
دانشیار، عضو هیأت علمی گروه سیستم ها و اقتصاد، موسسه عالی آموزش و پژوهش مدیریت و برنامه ریزی، تهران، ایران.
farhang@imps.ac.ir
علیرضا رشیدی کمیجان
دانشیار، عضو هیأت علمی گروه مهندسی صنایع، دانشکده فنی و مهندسی، واحد علوم و تحقیقات دانشگاه آزاد اسلامی، تهران ، ایران.
al.rashidi@iau.ac.ir
رویا سلطانی
استادیار، عضو هیأت علمی گروه مهندسی صنایع، دانشکده فنی و مهندسی، دانشگاه خاتم، تهران، ایران.
r.soltani@khatam.ac.ir
تاریخ دریافت: ۱۴۰۴/۰۸/۲۹ تاریخ پذیرش: ۱۴۰۴/۱۲/۰۳

چکیده
فرار مالیاتی یکی از مهم‌ترین چالش‌های نظام مالیاتی به شمار می‌رود و بخش قابل‌توجهی از آن ناشی از معاملات اشخاص وابسته، جابه‌جایی سود، و رفتارهای شبکه‌ای میان شرکت‌هاست. پیچیدگی ساختارهای مالکیتی، زنجیره‌های تراکنشی چندلایه و روابط پنهان بین شرکت‌ها باعث می‌شود روش‌های سنتی شناسایی فرار مالیاتی توان کافی برای تحلیل الگوهای سازمان‌یافته نداشته باشند. در پاسخ به این خلأ، پژوهش حاضر یک چارچوب گراف‌محور مبتنی بر کشف جامعه با الگوریتم لوین و شبکه عصبی گرافی GraphSAGE ارائه می‌کند تا بتواند ساختارهای پنهان تبانی، خوشه‌های غیرعادی و احتمال فرار مالیاتی در شبکه‌ای شامل بیش از ۱۶۲ هزار شرکت ایرانی را شناسایی کند. داده‌های پژوهش شامل ۳۳ شاخص مالی، بانکی و اطلاعات مالکیتی بوده و بیش از ۳۰۰ هزار رابطه تراکنشی و ساختاری میان شرکت‌ها استخراج شد. نتایج نشان می‌دهد مدل پیشنهادی به AUC معادل ۰،۰۹۵۷۹، دقت ۰،۰۸۹۰۵ و امتیاز F1 برابر ۰،۰۸۹۴۴ دست می‌یابد و قادر است جوامع پرریسک و الگوهای شبکه‌ای مؤثر در فرار مالیاتی را با تفکیک‌پذیری بالا شناسایی کند. همچنین تحلیل جامعه‌ای نشان می‌دهد بخش مهمی از روابط غیرعادی، در خوشه‌های متراکم شرکت‌های وابسته و گروه‌های اقتصادی بروز می‌کند. به‌طور کلی، یافته‌ها بیان می‌کند ترکیب کشف جامعه و GraphSAGE ابزاری کارآمد برای شناسایی الگوهای پیچیده فرار مالیاتی در داده‌های بزرگ است و می‌تواند به بهبود فرآیندهای تحلیل ریسک در نظام مالیاتی کمک کند.
واژه‌های کلیدی: تحلیل گراف، تشخیص اجتماع، شبکه های عصبی گراف، فرار مالیاتی، کشف تقلب مالیاتی، یادگیری استقرایی.

۱- مقدمه

تقلب و فرار مالیاتی از مهم‌ترین تهدیدها برای عدالت مالیاتی و کارایی درآمدهای عمومی محسوب می‌شوند. سازمان‌های بین‌المللی و پژوهش‌های آماری نشان می‌دهند که شکاف‌های مالیاتی و زیان‌های ناشی از فرار و اجتناب از مالیات در بسیاری از کشورها عدد قابل توجهی است و دولت‌ها به‌دنبال ابزارهای تحلیلی پیشرفته برای کاهش این شکاف‌اند(OECD ، ۲۰۲۳).

در ایران، الگوهای فرار مالیاتی غالباً بیش از آنکه صرفاً فردی باشند، شبکه‌ای و سازمان‌یافته‌اند؛ به‌ویژه معاملات اشخاص وابسته، جابه‌جایی سود درون‌گروهی و سازوکارهای مالکیتی پیچیده که امکان پنهان‌سازی جریان‌های سود و درآمد را فراهم می‌کنند. پژوهش‌های داخلی اخیر بر نقش تعیین‌کننده تعاملات گروهی در شکل‌گیری رفتارهای فرار مالیاتی تأکید کرده‌اند و مطالعه‌ای جدید نشان داد که بخش مهمی از روابط پرریسک در خوشه‌های متراکم شرکت‌های وابسته تمرکز دارد (احمدپور و همکاران، ۱۴۰۴) .

روش‌های سنتی تشخیص تقلب و فرار مالیاتی (قواعد، تحلیل‌های آماری ساده و مدل‌های داده‌کاوی غیرساختاریافته) در مواجهه با این پیچیدگی‌ها محدودیت‌هایی دارند؛ چرا که این روش‌ها معمولاً روابط بین‌مؤدیان را به‌صورت صِرفِ ویژگی‌های جداگانه تحلیل می‌کنند و از ساختار شبکه‌ای تبانی غافل می‌مانند. مرورهای قدیمی‌تر و پایه نشان می‌دهد که داده‌کاوی کلاسیک گرچه مفید است، اما در کشف الگوهای چندمرحله‌ای و گروهی ناکافی است(نگای و همکاران، ۲۰۱۱).

در سال‌های اخیر، شبکه‌های عصبی گراف[1] به‌عنوان رویکردی توانمند برای تحلیل داده‌های دارای ساختار رابطه‌ای معرفی شده‌اند؛ این مدل‌ها با مکانیزم انتقال پیام[2] قادرند هم‌زمان از ویژگی‌های گره‌ها و ساختار توپولوژیک گراف بهره ببرند و الگوهای پیچیده ارتباطی را آشکار سازند. مروری منظم بر پیشرفت‌های شبکه‌های عصبی گراف نشان می‌دهد که این خانواده از مدل‌ها در طیف گسترده‌ای از مسائل (از جمله تشخیص تقلب مالی) عملکرد چشمگیری داشته‌اند و راهکارهای جدیدی برای نگهداری مقیاس و تفسیرپذیری ارایه کرده‌اند (وو و همکاران، ۲۰۲۱) .

از میان معماری‌های شبکه‌های عصبی گراف، GraphSAGE با رویکرد «نمونه‌گیری و تجمیع محلی» توانسته محدودیت‌های مدل‌های ایستا را برطرف کند؛ GraphSAGE به‌صورت استقرایی تعبیه‌ای[3] تولید می‌کند که برای گره‌های جدید نیز قابل تعمیم است و بنابراین برای محیط‌های پویا مانند شبکه‌های مالیاتی مناسب‌تر است (همیلتون و همکاران، ۲۰۱۷). این قابلیت استقرایی هنگام مواجهه با ورود شرکت‌های تازه‌تأسیس یا تغییرات سریع در ساختار شبکه از اهمیت ویژه‌ای برخوردار است .

هم‌زمان، تحلیل جامعه‌ای[4] ابزاری مؤثر برای آشکارسازی خوشه‌های تبانی و گروه‌های هم‌عمل‌کننده است؛ الگوریتم لوین[5] به‌دلیل کارایی محاسباتی و کیفیت خوشه‌بندی در گراف‌های بزرگ، در بسیاری از مطالعات شبکه‌ای به‌کار رفته و می‌تواند مقدمات تفسیر نتایج مدل‌های گرافی را فراهم آورد (بلوندل و همکاران، ۲۰۰۸). ادغام نتایج کشف جامعه با فرایند تولید تعبیه‌ها، می‌تواند هم نویز را کاهش دهد و هم به تولید تعبیه‌هایی با خوانش جامعه‌محور کمک کند .

در حوزه تشخیص تقلب و فرار مالیاتی، مطالعات اخیر نشان می‌دهند که کاربرد شبکه‌های عصبی گراف و ساختارهای سلسله‌مراتبی (مثل مدل‌های توجه‌محور و ساختارهای ناهمگن) توانسته است دقت تشخیص را به‌طور قابل‌توجهی افزایش دهد و اشکال پیچیده تبانی را بهتر شناسایی کند(شی و همکاران، ۲۰۲۳). این پژوهش‌ها همچنین خلأهایی را درباره تفسیرپذیری، مقیاس‌پذیری و کاربرد در داده‌های واقعی چندصدهزار نودی برجسته کرده‌اند، که برای کاربرد در نظام‌های مالیاتی ملی اهمیت دارد.

با توجه به مراتب فوق، پژوهش حاضر با هدف پر کردن این شکاف‌ها چارچوبی ترکیبی ارائه می‌دهد که ابتدا با الگوریتم لوین جوامع پنهان را استخراج می‌کند و سپس با استفاده از GraphSAGE تعبیه‌های استقرایی تولید نموده و در نهایت روابط پرریسک (احتمال فرار مالیاتی) را پیش‌بینی می‌نماید. فرایند کلی مدل پیشنهادی در شکل شماره ۱ نمایش داده شده است .

[1] Graph Neural Network- GNN
[2] message passing
[3] embedding
[4] Community detection
[5] Louvain

شکل شماره ۱: فرایند کلی مدل پیشنهادی

نوآوری اصلی این مطالعه عبارت است از:

۱) تلفیق کشف جامعه لوین با یادگیری استقرایی GraphSAGEبرای تحلیل شبکه‌های مالیاتی بزرگ
۲) ارائه چارچوب گراف‌محور مقیاس‌پذیر که قابلیت تعمیم به گره‌های جدید و داده‌های در حال گسترش را دارد
۳) ارزیابی مدل روی یک شبکه واقعی با بیش از ۱۶۲ هزار شرکت و ۳۰۰ هزار رابطه که زمینه‌ای واقعی و بزرگ‌مقیاس برای تشخیص الگوهای فرار مالیاتی فراهم می‌کند.
۴) بهبود دقت تشخیص نسبت به مدل تعبیه‌محور پژوهش قبلی (پژوهش پذیرش یافته در مجله مدیریت و بهره‌وری، ۱۴۰۳) و سایر مدل‌های پایه.

ساختار مقاله به این شرح است: پس از این مقدمه، بخش دوم به مبانی نظری، بخش سوم به مرور ادبیات و پیشینه‌های پژوهشی مرتبط در حوزه‌های تشخیص تقلب مالیاتی، شبکه‌های عصبی گرافی و کشف جامعه می‌پردازد. در بخش چهارم، روش‌شناسی پژوهش شامل مجموعه‌داده، مراحل پیش‌پردازش، معماری پیشنهادی و جزئیات آموزش مدل ارائه می‌شود. بخش پنجم به تحلیل یافته‌ها و ارائه نتایج کمی و کیفی حاصل از ارزیابی مدل اختصاص دارد. در نهایت، در بخش ششم، نتیجه‌گیری و پیشنهادهایی برای پژوهش‌های آتی مطرح می‌گردد.

۲. مبانی نظری پژوهش

۲-۱ شبکه‌های عصبی گرافی

شبکه‌های عصبی گرافی به عنوان زیرشاخه‌ای از یادگیری ماشین، امکان یادگیری مستقیم از داده‌های گرافی را فراهم می‌آورند. برخلاف داده‌های جدولی یا تصویری که ساختار ثابتی دارند، گراف‌ها ساختار غیر اقلیدسی و نامنظمی را نمایش می‌دهند که مدل‌سازی آن با روش‌های سنتی دشوار است. هسته اصلی شبکه‌های عصبی گراف بر پایه مکانیزم انتشار پیام استوار است که طی آن هر گره (نود) اطلاعات خود را با همسایگانش به اشتراک می‌گذارد و بازنمایی غنی‌تری از ویژگی‌های ساختاری و محتوایی به دست می‌آورد.

فرآیند یادگیری در شبکه‌های عصبی گراف را می‌توان در هر لایه k به صورت رابطه شماره ۱ نمایش داد:

$$h_v^k = \phi\left(h_v^{k-1},\ \oplus_{\{u \in N(v)\}} \psi\left(h_v^{k-1}, h_u^{k-1}, e_{vu}\right)\right)$$

رابطه شماره ۱

که در آن:

$h_v^{(k)}$:بردار بازنمایی گره v در لایهk ، $h_v^{(k-1)}$: بازنمایی گره v در لایه قبلی، $h_u^{(k-1)}$: بازنمایی گره همسایه u در لایه قبلی ، N(v): مجموعه همسایگان گرهv ، e_{uv}: ویژگی یال میان گره v و u، ϕ: تابع پیام‌رسانی[1]، ϕ :تابع به‌روزرسانی[2]و $\oplus$ عملگر تجمیع[3]، مانند میانگین یا بیشینه

این فرمول پایه‌ای‌ترین شکل انتشار پیام است و اکثر معماری‌های شبکه‌های عصبی گراف نسخه‌ای از آن را به‌کار می‌گیرند.

۲-۲ معماریGraphSAGE

GraphSAGEیک معماری استقرایی است؛ یعنی می‌تواند برای گره‌های جدیدی که در زمان آموزش دیده نشده‌اند نیز بازنمایی تولید کند. این ویژگی برای تحلیل شبکه‌های مالیاتی که داده‌های جدید دائماً اضافه می‌شوند، اهمیت عملی دارد شکل شماره ۲ این معماری را به‌صورت شماتیک نمایش می‌دهد.

۲-۲-۱ مکانیزم نمونه‌گیری[4]

برای کنترل پیچیدگی محاسباتی، از هر گره تنها تعداد ثابتی از همسایگان برداشت می‌شود:

$$N_{k(v)} = SAMPLE(N(v), S_k)$$ رابطه شماره ۲

که در آن :

$N_{k(v)}$: مجموعه نمونه‌گیری‌شده از همسایگان گره v در لایهk ، N(v): همه همسایگان گرهv ، S_k: تعداد نمونه‌برداری از همسایگان در لایهk و SAMPLE تابع نمونه‌گیری تصادفی یا کنترل‌شده می‌باشد.

[1] Message Function
[2] Update Function
[3] Aggregation Operator
[4] Sampling

۲-۲-۲ توابع تجمیع[1]

الف) تجمیع میانگین[2]

رابطه شماره ۳

$$h_{N(v)}^{k} = \left(\frac{1}{|N(v)|}\right) * \ \Sigma_{\{u \in N(v)\}} h_u^{k-1}$$

$h_{N(v)}^{(k)}$: بردار ویژگی تجمیع‌شده برای همسایگان گره v در لایه k، $|N(v)|$: تعداد همسایگان گرهv ، Σ: جمع روی همسایگان و $h_u^{(k-1)}$: ویژگی گره u در لایه قبلی است.

ب) تجمیع LSTM

رابطه شماره ۴

$$h_{N(v)}^{k} = LSTM(\{ h_u^{k-1} \ | u \in N(v) \})$$

$h_{N(v)}^{(k)}$: خروجی LSTM برای همسایگان گرهv ، LSTM : شبکه‌ای برای یادگیری توالی ویژگی‌های همسایگان است و $\{ h_u^{(k-1)} \}$ مجموعه بردارهای ویژگی همسایگان در لایه قبلی می‌باشد.

ج) تجمیع پولینگ[3]

رابطه شماره ۵

$$h_{N(v)}^{k} = \max\{ \sigma(W_{pool} * h_u^{k-1} + b), \quad \forall u \in N(v)\}$$

$h_{N(v)}^{(k)}$: بردار ویژگی تجمیع‌شده با پولینگ برای گرهv ، max : عملگر بیشینه‌گیری روی همسایگان، σ: تابع فعال‌سازی مانند ReLU، W_{pool}: ماتریس وزن‌های لایه پولینگ، b: بردار بایاس و $h_u^{(k-1)}$ بردار ویژگی همسایه u در لایه قبلی است.

۲-۲-۳ به‌روزرسانی بازنمایی گره

پس از تجمیع، ویژه‌برداری گره اصلی با الحاق خود گره و ویژگی همسایگان به‌روزرسانی می‌شود:

رابطه شماره ۶

$$h_v^{k} = \sigma(W^k * CONCAT(h_v^{k-1}, h_{N(v)}^{k}))$$

$h_v^{(k)}$: بازنمایی جدید گره v در لایهk ، σ: تابع فعال‌سازی مثلاً ReLUیا سیگموید[4]، $W^{(k)}$: ماتریس وزن قابل یادگیری در لایه k، CONCAT: عملگر الحاق بردارها، $h_v^{(k-1)}$: ویژگی قبلی گره vو $h_{N(v)}^{(k)}$ ویژگی تجمیع‌شده همسایگان است.

GraphSAGE به دلیل ساختار نمونه‌برداری، قابلیت کار در گراف‌های بسیار بزرگ را دارد و بازنمایی‌های تولیدشده قابل استفاده برای گره‌های دیده‌نشده هستند. این ویژگی در ردیابی رفتارهای مالی متغیر و پویا کاربرد مستقیم دارد.

۳. پیشینه پژوهش

در سال‌های اخیر، پیچیده‌تر شدن ساختارهای معاملاتی و افزایش حجم تعاملات مالی باعث شد روش‌های کلاسیک یادگیری ماشین در شناسایی الگوهای پنهان فرار مالیاتی با محدودیت‌های جدی مواجه شوند، زیرا این روش‌ها اغلب روابط میان مؤدیان را به‌صورت مجزا و جدولی تحلیل می‌کنند و ساختارهای شبکه‌ای پیچیده را نادیده می‌گیرند. در مقابل، تحلیل گراف و یادگیری عمیق گرافی با نمایاندن داده‌ها به شکل گراف — که گره‌ها نمایانگر نهادها و لبه‌ها نمایانگر روابط مالی هستند — توانایی به‌مراتب بالاتری در استخراج الگوهای مخفی، تبانی‌های چندمرحله‌ای و ویژگی‌های ساختاری دارد (وانگ و همکاران، ۲۰۱۶؛ ژانگ و ژو، ۲۰۱۷).

در نخستین گام، رویکردهای تعبیه‌سازی گراف مانند گشت عمقی[5] مطرح شدند که با استفاده از گشت‌وگذارهای تصادفی، گره‌های شبکه را به بردارهایی در فضای کم‌بعد تبدیل می‌کردند. کمی بعد، مدل بازنمایی گره به بردار[6] با معرفی راهبردهای انعطاف‌پذیرتر در گشت‌وگذار از جمله حرکت اکتشافی یا تمرکز روی همسایگی نزدیک توانست شباهت‌های ساختاری میان مؤدیان را بهتر ثبت کند. این روش‌ها امکان استفاده از طبقه‌بندهای کلاسیک را فراهم می‌کردند، اما یک ضعف مهم داشتند: تعبیه‌های آن‌ها استقرایی نبود و برای گره‌های جدید که در سامانه‌های مالیاتی دائماً ظاهر می‌شوند قابلیت تعمیم نداشتند (پرو و همکاران، ۲۰۱۶؛ گروور و لسکوک، ۲۰۱۶).

محدودیت‌های یادشده، راه را برای شکل‌گیری نسل جدید مدل‌های گرافی یعنی شبکه‌های عصبی گراف هموار کرد. نقطه تمایز این شبکه‌ها، استفاده از سازوکار «انتقال پیام» است که طی آن، هر گره اطلاعات خود را با همسایگان به اشتراک می‌گذارد و یک نمایش غنی‌تر از ساختار محلی شبکه ایجاد می‌شود. یکی از اولین معماری‌های مهم این حوزه، شبکه کانولوشن گرافی[7] بود که توانست روابط چندمرحله‌ای میان گره‌ها را استخراج کند، اما در وزن‌دهی به روابط مختلف قادر به تمایز نبود و همه یال‌ها را یکسان فرض می‌کرد (کیپف و ویلینگ، ۲۰۱۷).

[1] Aggregation Functions
[2] Mean Aggregator
[3] Pooling Aggregator
[4] sigmoid
[5] Deep Walk
[6] Node2Vec
[7] Graph Convolutional Network-GCN

برای رفع این ضعف، شبکه توجه گرافی[1] معرفی شد که با استفاده از سازوکار توجه، به هر همسایه وزن متفاوتی اختصاص می‌داد. در سیستم‌های مالیاتی، اهمیت روابط یکسان نیست؛ یک تراکنش خاص یا مالکیت مشترک می‌تواند نقش کلیدی در یک زنجیره تبانی ایفا کند. به همین دلیل، شبکه‌های توجه گرافی از لحاظ نظری توانایی بیشتری در شناسایی روابط معنادار دارد (ولاچیس و همکاران، ۲۰۱۸).

گام بعدی تکامل مدل‌های گرافی با ظهور GraphSAGE رقم خورد. این مدل یک نوآوری اساسی داشت: استقرایی بودن . GraphSAGEبه‌جای آن‌که کل گراف را به‌طور کامل پردازش کند، از همسایگان نمونه‌گیری می‌کند و با استفاده از توابع تجمیع مختلف مانند میانگین، LSTMو تجمیع مبتنی بر بیشینه ؛ قابلیت تعمیم به گره‌های جدید را فراهم می‌کند. همین ویژگی باعث شد GraphSAGE به یکی از مناسب‌ترین مدل‌ها برای تحلیل شبکه‌های مالیاتی تبدیل شود؛ زیرا در چنین محیط‌هایی، ظهور شرکت‌ها و روابط جدید امری طبیعی است (همیلتون، ییننگ و لسکوک، ۲۰۱۷).

در این مسیر، پژوهش‌های متعددی تلاش کردند تا مدل‌های گرافی را مستقیماً در تشخیص تقلب مالی به‌کار گیرند. یکی از پیشرفت‌های اخیر، چارچوب FinGuard-GNN است که از شبکه‌های عصبی گراف پویا برای تشخیص تقلب مالی بهره می‌برد و با به‌کارگیری استراتژی‌های نوین در تبدیل ویژگی‌ها و انتشار ریسک، عملکرد خود را نسبت به روش‌های سنتی ارتقا داده است (هوآنگ، ۲۰۲۵). پژوهشی دیگر نشان داده است که مدل‌های گرافی می‌توانند معاملات را به صورت گراف‌های ناهمگن مدل کنند، به‌طوری‌که گره‌ها انواع مختلف موجودیت‌ها را نشان می‌دهند و لبه‌ها روابط چندوجهی بین آن‌ها را ضبط می‌کنند، که باعث می‌شود الگوهای شبکه‌ای تقلب پیچیده‌تر قابل شناسایی شوند (تاکاهی‌شی و همکاران، ۲۰۲۵).

چالش‌های زمانی و پویا بودن روابط مالی نیز توجه پژوهشگران را به خود جلب کرده است. به‌عنوان مثال، چارچوب‌های گراف عصبی مانند TeMP-TraG با در نظر گرفتن دینامیک زمانی تراکنش‌ها، پیام‌رسانی را اولویت‌بندی می‌کنند تا الگوهای حساس به زمان را بهتر آشکار سازند و بهبود عملکرد نسبت به مدل‌های گرافی سنتی نشان دهند (گونویه و همکاران، ۲۰۲۵). همچنین، مدل‌های نوین مانند DynBERG با ترکیب معماری‌های ترنسفورمر مبتنی بر گراف و اجزای زمانی قادر به تحلیل تغییرات ساختاری در شبکه‌های تراکنشی هستند که در داده‌های رمزارزها و بازارهای پویا اهمیت دارد (کولکارنی و چاندرا، ۲۰۲۵).

در زمینه تشخیص تقلب بلادرنگ، پژوهش‌هایی مطرح شده‌اند که از گراف عصبی تطبیقی برای تشخیص تقلب در زمان واقعی با حفظ حریم خصوصی داده‌ها استفاده می‌کنند (رحماتی، ۲۰۲۵). این رویکردها نشان می‌دهند که ترکیب یادگیری گرافی با روش‌های نوین داده‌محور می‌تواند هم دقت و هم کارایی را در محیط‌های واقعی ارتقا دهد.

در کنار این تحولات در ادبیات بین‌المللی، در ایران نیز پژوهش‌هایی در حوزه حسابداری و حسابرسی مدیریت به بررسی تقلب، گزارشگری مالی و ریسک‌های مرتبط پرداخته‌اند. برای مثال، مرادی و همکاران (۱۴۰۲) در «فصلنامه دانش حسابداری و حسابرسی مدیریت» نشان دادند که گرایش مدیران به ریسک و انتظارات رفتاری آن‌ها می‌تواند نقش معناداری در بروز گزارشگری مالی تقلب‌آمیز داشته باشد. این نتایج بیانگر آن است که تقلب مالی صرفاً ناشی از متغیرهای عددی نیست، بلکه عوامل رفتاری و مدیریتی نیز در شکل‌گیری آن مؤثرند (مرادی و همکاران، ۱۴۰۲).

همچنین حیرانی و همکاران (۱۴۰۰) با تمرکز بر ریسک حسابرسی مالیاتی، شاخصه‌های مؤثر بر این ریسک را شناسایی کردند و نشان دادند که ساختار اطلاعاتی، پیچیدگی معاملات و ضعف در شفافیت می‌تواند احتمال بروز خطا یا تقلب مالیاتی را افزایش دهد. یافته‌های این پژوهش اهمیت تحلیل ساختارهای پیچیده مالی و استفاده از رویکردهای داده‌محور را در نظام حسابرسی مالیاتی برجسته می‌کند (حیرانی و همکاران، ۱۴۰۰).

در پژوهشی دیگر، مددی‌زاده و همکاران (۱۴۰۱) تأثیر استراتژی‌های مالیاتی جسورانه و تقلب حسابداری را بر خوانایی گزارشگری مالی بررسی کردند و نشان دادند که رفتارهای پرریسک مالیاتی و تقلب‌آمیز می‌تواند کیفیت و شفافیت گزارشگری مالی را تضعیف کند. این نتایج حاکی از آن است که تقلب و فرار مالیاتی اغلب در بستری از روابط و تصمیمات به‌هم‌پیوسته شکل می‌گیرد که تحلیل آن‌ها نیازمند ابزارهای پیشرفته‌تر از روش‌های سنتی است (مددی‌زاده و همکاران، ۱۴۰۱).

از سوی دیگر، در کنار پژوهش‌های بین‌المللی در حوزه گراف کاوی، در پژوهش‌های داخلی نیز به بررسی نقش تحلیل گراف در مسائل مالیاتی پرداخته است. برای نمونه، مطالعه امین احمدپور، سیده محبوبه جعفری و فاطمه صراف (۱۴۰۴) با ترکیب گراف‌کاوی و شبکه‌های عصبی عمیق به مدل‌سازی فرار

[1] Graph Attention Network

مالیاتی معاملات اشخاص وابسته پرداخته و نشان داده است که تحلیل ساختار شبکه‌ای می‌تواند الگوهای پنهان رفتارهای هماهنگ را بهتر از روش‌های سنتی آشکار سازد (احمدپور و همکاران، ۱۴۰۴). علاوه بر این، پژوهشی مانند زارع بهنمیری و همکاران (۱۴۰۱) عوامل مؤثر بر فرار مالیاتی را با استفاده از روش‌های تصمیم‌گیری چندمعیاره رتبه‌بندی کرده است، که اهمیت درک بهتر سازوکارهای پیچیده اقتصادی و نیاز به روش‌های جدید تحلیل داده را برجسته می‌کند (زارع بهنمیری و همکاران، ۱۴۰۱).

در مجموع، مرور ادبیات نشان می‌دهد که روند تکامل مدل‌های گرافی تا پیش از ظهور مدل‌های زمانی، از تعبیه‌سازی‌های ساده به سمت شبکه‌های عصبی گرافی پیش رفته است و GraphSAGE نقطه بلوغ این نسل از مدل‌ها محسوب می‌شود؛ مدلی که میان دقت، مقیاس‌پذیری و قابلیت تعمیم تعادل ایجاد کرده و برای تحلیل شبکه‌های مالیاتی بزرگ مناسب است.

جدول شماره ۱: مقایسه پژوهش‌های پیشین مرتبط با کشف تقلب و فرار مالیاتی مبتنی بر گراف

پژوهش	نوع داده / حوزه	روش / مدل	نقاط قوت	محدودیت‌ها
گروور و لسکوک (۲۰۱۶) – Node2Vec	شبکه تراکنش و روابط سازمانی	تعبیه‌سازی گراف با گشت‌وگذار هدایت‌شده	استخراج روابط ساختاری عمیق؛ مناسب برای کشف الگوهای تبانی	استقرایی نبودن؛ نیاز به پردازش کامل گراف پس از ورود گره جدید
پرو و همکاران (۲۰۱۶) – Walk Deep	شبکه‌های تعامل مالی	تعبیه‌سازی مبتنی بر گشت‌وگذار تصادفی	ساده، سریع، مناسب برای کشف خوشه‌های پنهان	ناتوان در وزن‌دهی به روابط و لایه‌بندی ساختار
کیپف و ویلینگ (۲۰۱۷) – GCN	شبکه تراکنش مالی و مالکیت	شبکه عصبی کانولوشن گرافی	توانایی استخراج روابط دو-سه‌مرحله‌ای؛ دقت مناسب در شناسایی رفتارهای مشابه	وزن‌دهی یکنواخت به یال‌ها؛ ناتوانی در تشخیص روابط مهم‌تر
ولاچیس و همکاران (۲۰۱۸) – GAT	شبکه‌های مالی با روابط ناهمگن	شبکه توجه گرافی	توانایی تمایز بین روابط ضعیف و قوی؛ مناسب برای تحلیل تبانی	هزینه محاسباتی بالا؛ حساسیت به پارامترهای توجه
همیلتون، یینگ و لسکوک (۲۰۱۷) – GraphSAGE	شبکه تراکنش مؤدیان، شرکت‌ها و شرکا	مدل استقرایی با تجمیع همسایگی	قابلیت تعمیم به مؤدیان جدید؛ مناسب برای شبکه‌های بزرگ مالیاتی	وابستگی به انتخاب نوع تجمیع؛ کاهش دقت در شبکه‌های بسیار پراکنده
لی و همکاران (۲۰۱۶)	شبکه مالکیت و صورت‌های مالی	ترکیب ویژگی‌های مالی + ساختار گراف	افزایش دقت کشف تبانی نسبت به روش‌های جدولی	عدم توجه به اهمیت نسبی انواع روابط
شی و ژائو (۲۰۱۷)	گراف روابط تجاری	خوشه‌بندی مبتنی بر چگالی و شاخص‌های گراف	کشف خوشه‌های تبانی و حلقه‌های معاملاتی	وابستگی شدید به کیفیت و تراکم داده‌ها
پژوهش	**نوع داده / حوزه**	**روش / مدل**	**نقاط قوت**	**محدودیت‌ها**
احمدپور و همکاران (۱۴۰۴) – معاملات اشخاص وابسته	داده‌های واقعی شرکت‌های ایرانی	تحلیل شبکه + منطق فازی	شناسایی ساختارهای غیرعادی در معاملات وابسته	فاقد مدل یادگیری عمیق؛ عدم پیش‌بینی الگوهای پنهان
زیا و همکاران (۲۰۲۴)	شبکه تراکنش مالی-کارت‌های اعتباری	استخراج ویژگی‌های شبکه + SVM	کارایی بالا در داده‌های کوچک	عملکرد ضعیف در ساختارهای غیرخطی و پیچیده
مطالعه قبلی پژوهشگر (پذیرش‌شده و چاپ در فروردین ۱۴۰۵)	شبکه خرید و فروش شرکت‌ها با داده‌های واقعی	بازنمایی گراف + طبقه‌بندی	تشخیص حلقه‌های فرار مالیاتی؛ تحلیل ساختارهای تبانی	محدودیت در پیش‌بینی مؤدیان جدید؛ فقدان لایه تعمیم‌یافته

همانطور که از جدول شماره ۱ پیداست از تعبیه‌سازی‌های اولیه مانند گشت عمقی و بازنمایی گره به بردار تا ظهور معماری‌های اولیه شبکه‌های عصبی گراف مثل شبکه گراف پیچشی[1]، شبکه گراف توجه و نهایتا GraphSAGE چند شکاف کلیدی مشخص می‌شود:

۱- نبود مدل‌های استقرایی در تحقیقات مالیاتی ایران

اگرچه GraphSAGE در سطح بین‌المللی معرفی شده، اما در پژوهش‌های ایرانی هنوز تأکید بر شاخص‌های شبکه یا روش‌های تعبیه‌سازی ساده است. بنابراین تشخیص مؤدیان جدید با ساختارهای نوظهور تقریباً در هیچ مطالعه داخلی انجام نشده است.

۲- عدم استفاده از ترکیب ساختار گراف و ویژگی‌های مالی واقعی کسب‌وکار

عدم بهره‌گیری از ترکیب هم‌زمان ساختار گراف روابط اقتصادی و ویژگی‌های مالی واقعی کسب‌وکارها یکی از محدودیت‌های اصلی پژوهش‌های پیشین در حوزه شناسایی تقلب و فرار مالیاتی محسوب می‌شود. بخش قابل‌توجهی از مطالعات، تمرکز خود را صرفاً بر تحلیل شبکه و الگوهای ارتباطی میان اشخاص یا بنگاه‌ها قرار داده‌اند، در حالی که گروه دیگری از پژوهش‌ها تنها بر شاخص‌ها و متغیرهای مالی متکی بوده‌اند. این جداسازی رویکردها، موجب شده است الگوهای تبانی که ذاتاً حاصل تعامل هم‌زمان روابط شبکه‌ای و رفتارهای مالی هستند، به‌طور کامل شناسایی نشوند. ازاین‌رو، فقدان چارچوب‌هایی که بتوانند این دو بُعد اطلاعاتی را به‌صورت یکپارچه مدل‌سازی کنند، همچنان به‌عنوان یک خلأ مهم در ادبیات پژوهش مطرح است.

۳- فقدان تحلیل جهت‌دار بودن روابط

روابط مالیاتی (جریان پول، خرید/فروش، مالکیت) جهت‌دار هستند، اما بیشتر مدل‌های استفاده‌شده آن‌ها را نامتقارن یا بدون جهت فرض کرده‌اند که باعث از دست رفتن الگوهای مهم می‌شود.

۴- ناکارآمدی روش‌های قبلی در شبکه‌های بزرگ و پویا

گرچه GraphSAGE برای مقیاس‌پذیری طراحی شده، ولی هیچ پژوهش داخلی تاکنون از آن در شبکه‌های بزرگ شرکت‌های ایرانی استفاده نکرده است.

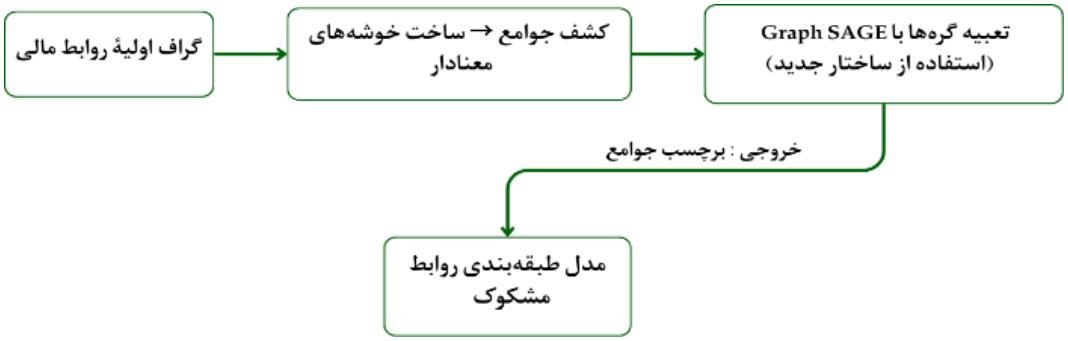
۵-کمبود تمرکز بر کشف حلقه‌های تبانی و جوامع پنهان

بیشتر مطالعات فقط بر طبقه‌بندی گره‌ها تمرکز دارند، در حالی‌که پیدایش حلقه‌ها[2] و جوامع تبانی ویژگی اصلی فرار مالیاتی است. ترکیب GraphSAGE و تشخیص اجتماعات تاکنون در ادبیات مالیاتی بررسی نشده است.

۳. روش‌شناسی پژوهش

۳-۱ طرح کلی پژوهش

هدف این پژوهش توسعه‌ی یک چارچوب تحلیل شبکه‌ای برای شناسایی الگوهای پنهان فرار مالیاتی در شبکه معاملات شرکت‌هاست. بدین‌منظور، ابتدا ساختار شبکه‌ای تراکنش‌ها بازسازی شده، سپس الگوهای جامعه‌ای[3] و حلقه‌های معاملاتی شناسایی می‌شوند. در گام بعد، از مدل GraphSAGE برای یادگیری نمایش‌های استقرایی از مؤدیان استفاده می‌شود تا بتوان مؤدیان جدید و روابط نوظهور را نیز با دقت بالا طبقه‌بندی کرد. ترکیب هم‌زمان «کشف جامعه» و «GraphSAGE» باعث افزایش دقت در تشخیص رفتارهای تبانی می‌شود، زیرا هم اطلاعات ساختار کلان شبکه و هم الگوهای محلی گره‌ها لحاظ می‌گردد. مدل مفهومی در شکل شماره ۲ آمده است.



شکل شماره ۲: مدل مفهومی ارتباط Louvain و GraphSAGE

[1] GCN
[2] loops
[3] Communities

۳-۲ بازسازی شبکه مالیاتی

۳-۲-۱ تعریف گره‌ها و یال‌ها

در این پژوهش، شبکه مالیاتی شرکت‌ها به‌صورت یک گراف جهت‌دار G=(V,E) تعریف می‌شود که در آن:

V مجموعه شرکت‌ها (مؤدیان حقیقی/حقوقی)، E مجموعه روابط معاملاتی (خرید، فروش، صدور فاکتور، قرارداد) و جهت یال از فروشنده به خریدار است.

محاسبه وزن یال از شرکت i به شرکت j :

رابطه شماره ۷

$$w_{ij} = \log(1 + \text{Amount}_{ij})$$

استفاده از لگاریتم در رابطه شماره ۷ باعث کاهش تأثیر مقادیر بسیار بزرگ و کنترل نوسانات شدید داده‌های مالی می‌شود.

۳-۲-۲ ویژگی‌های گره‌ها و یال‌ها

هر شرکت با مجموعه‌ای از ویژگی‌های مالی و ساختاری توصیف می‌شود که در پنج دسته اصلی قرار می‌گیرند:

فراداده شرکت (نوع صنعت، بخش اداری، نوع شرکت)

معیارهای تراکنش بانکی (کل دریافتی‌ها، کل بدهی‌ها، مانده بانکی)

معیارهای مالیاتی (درآمد اعلام شده، مالیات تعلق گرفته، مالیات پرداختی)

جریمه‌ها و معافیت‌ها (جریمه معوق، جریمه پرداختی، معافیت‌ها)

صورت‌های سود و زیان (سود انباشته، سود خاص)

روابط و یال‌ها: یال‌ها بین شرکت‌هایی ایجاد شده‌اند که دارای روابط تجاری، اشتراک اعضای هیئت مدیره، یا ارتباطات مالی دیگر هستند.

۳-۳ شناسایی جوامع پنهان[1]

برای شناسایی زیرشبکه‌های مشکوک از الگوریتم لوین استفاده می‌شود . بر اساس بیشینه‌سازی شاخص ماژولاریتی[2] عمل می‌کند:

رابطه شماره ۸

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j)$$

در رابطه شماره ۸، A_{ij} وزن یال‌ها، K_i و K_j مجموع وزن یال‌های متصل به گره i و j ، m مجموع وزن کل یال‌ها و δ تابع نشانگر تعلق به جامعه که چنانچه دو گره i و j به یک جامعه تعلق داشته باشند ۱ و در غیر اینصورت ۰ است.

الگوریتم لوین در این پژوهش، با شناسایی خوشه‌های متراکم از گره‌ها (جوامع)، بستری برای کشف الگوهای مشکوک فراهم می‌کند. از مهم‌ترین کاربردهای آن می‌توان به تشخیص حلقه‌های تبانی درون‌گروهی و بین‌گروهی، و همچنین شناسایی اجتماعاتی با حجم غیرعادی و متمرکز تراکنش‌ها اشاره کرد. همچنین، یک ویژگی جدید با عنوان «شناسه جامعه» از خروجی این الگوریتم استخراج و به عنوان ورودی در مدل GraphSAGE مورد استفاده قرار گرفت تا توانایی مدل در یادگیری الگوهای رفتاری هر جامعه بهبود یابد.

۳-۳-۱ ادغام جامعه با ویژگی مالی

رابطه شماره ۹

$$x_v^{final} = CONCAT\left(x_v^{financial}, \quad OneHot(C(v))\right)$$

x_v^{final}: بردار ویژگی نهایی گره v پس از ادغام اطلاعات

$x_v^{financial}$: ویژگی‌های مالی گره v

C(v): شناسه جامعه‌ای که گره v به آن تعلق دارد

OneHot(C(v)) : نمایش وان هات جامعه گره v

۳-۴ مدل یادگیری استقرایی GraphSAGE

مدل GraphSAGE با هدف یادگیری بازنمایی‌های برداری (تعبیه‌ها) از گره‌ها طراحی شده است. ویژگی برجسته و کلیدی این مدل، قابلیت یادگیری استقرایی آن می‌باشد که در این پژوهش، امکان طبقه‌بندی شرکت‌های جدید و دیده‌نشده را بدون نیاز به فرآیند پرهزینه و زمان‌بر آموزش مجدد مدل فراهم می‌سازد.

۳-۴-۱ سازوکار اصلی GraphSAGE

معماری GraphSAGE پیاده‌سازی شده در این پژوهش شامل اجزای مندرج در جدول شماره ۲ می‌باشد.

[1] Community Detection

[2] Modularity

جدول شماره ۲: مشخصات معماری GraphSAGE پیاده‌سازی‌شده در پژوهش

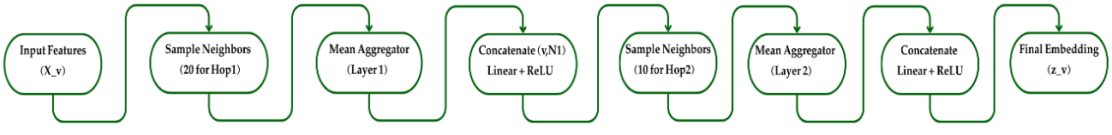
مؤلفه	پیکربندی و مقادیر	توضیحات تکمیلی
لایه ورودی	• ابعاد: ۳۳ ویژگی	مطابق با تعداد ویژگی‌های مالی و ساختاری هر گره
	• پیش‌پردازش: نرمال‌سازی Z-Score	
لایه‌های پنهان	• تعداد لایه‌ها: ۲	-
	• واحدهای هر لایه: ۲۰	
	• تابع فعال‌ساز: ReLU	
	• نرخ حذف تصادفی: ۰.۳	
مکانیزم نمونه‌گیری	• پرش اول: ۲۰ همسایه	نمونه‌گیری تصادفی از همسایگان در هر عمق
	• پرش دوم: ۱۰ همسایه	
	• رشد میدان دریافت: ۲۰ ⟶ ۲۰۰ گره	
مکانیزم تجمیع	• تابع تجمیع: میانگین	-
	• ترکیب: الحاق اطلاعات گره مرکزی و همسایگان	
مؤلفه	پیکربندی و مقادیر	توضیحات تکمیلی
لایه خروجی	• ابعاد خروجی: ۱	برای پیش‌بینی وجود یال بین جفت گره‌ها
	• محاسبه شباهت: ضرب داخلی	
پارامترهای آموزش	• تابع زیان: آنتروپی متقاطع	-
	• بهینه‌ساز: Adam	
	• نرخ یادگیری: ۰.۰۰۱	
	• دوره‌های آموزش: ۲۰	

در شکل شماره ۳ شماتیک معماری GraphSAGE به‌کاررفته در پژوهش نمایش داده شده است. این نمودار، فرآیند تولید تعبیه نهایی برای یک گره (v) در معماری GraphSAGE را نشان می‌دهد. فرآیند به‌صورت گام‌به‌گام و به‌شرح زیر است:

برای ورودی ویژگی‌های اولیه گره (x_v) به‌عنوان نقطه شروع استفاده می‌شود. در پرش اول از بین همسایگان مستقیم گره v، تعداد ۲۰ همسایه به‌صورت تصادفی نمونه‌گیری می‌شوند. اطلاعات این همسایگان با استفاده از تابع تجمیعگر میانگین[1] در لایه اول ادغام می‌شود. نمایندگی تجمیع‌شده همسایگان (N1)با نمایندگی گره مرکزی (v) الحاق[2] می‌شود. نتیجه حاصل از یک تبدیل خطی[3] و سپس تابع فعال‌ساز ReLU عبور داده می‌شود تا بازنمایی جدید گره ($h_v^{(1)}$) به‌دست آید. سپس در پرش دوم، این فرآیند برای همسایگان دورتر تکرار می‌شود. از بین همسایگان گره‌های نمونه‌گیری‌شده در پرش اول، تعداد ۱۰ همسایه نمونه‌گیری می‌شود. اطلاعات آن‌ها با تجمیع‌گر میانگین در لایه دوم ادغام (N2) و مجدداً با بازنمایی گره مرکزی (h_v) ($^{(1)}$) الحاق می‌شود. خروجی پس از عبور از تبدیل خطی و ReLU، بازنمایی نهایی (z_v) را تشکیل می‌دهد.

خروجی این معماری تعبیه نهایی گره (z_v) که حاوی اطلاعاتِ همسایگان تا عمق ۲ و ویژگی‌های ذاتی خود گره است، برای انجام وظیفه پایانی (مانند پیش‌بینی ارتباط) مورد استفاده قرار می‌گیرد.

این معماری با درنظرگیری سلسله‌مراتب همسایگی، امکان یادگیری بازنمایی‌های غنی و مقیاس‌پذیر را برای گراف‌های بزرگ فراهم می‌کند.



شکل شماره ۳: معماری داخلی GraphSAGE دو لایه‌ای

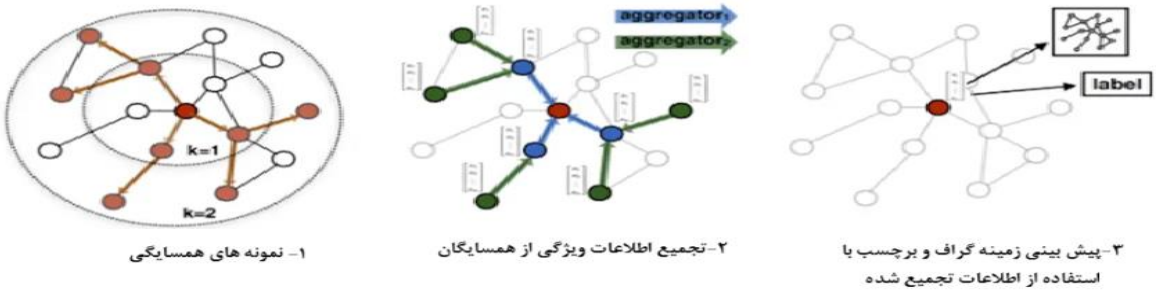
[1] Mean Aggregator
[2] Concatenate
[3] Linear

این فرآیند شامل سه مکانیزم اصلی زیر است که در شکل شماره ۴ نیز نمای شماتیک آن آمده است:

۱ . نمونه‌گیری همسایگی: به‌جای استفاده از تمام همسایگان که از نظر محاسباتی پرهزینه است، از هر گره تعداد ثابتی از همسایگان به‌صورت تصادفی نمونه‌گیری می‌شود. این نمونه‌گیری تا عمق K (در اینجا ۲) انجام می‌شود تا اطلاعات از همسایگان دورتر نیز جمع‌آوری گردد.

۲ . توابع تجمیع: این توابع، اطلاعات همسایگان نمونه‌گیری‌شده را در هر لایه ادغام می‌کنند. در این پژوهش از تابع تجمیع میانگین استفاده شده که با میانگین‌گیری از ویژگی‌های همسایگان، تعادل مناسبی بین سرعت و دقت برقرار می‌کند.

۳ . فرآیند تولید تعبیه: فرآیند یادگیری به‌صورت لایه‌به‌لایه و بازگشتی انجام می‌شود و در نهایت پیش‌بینی تعبیه نهایی گره به‌دست می‌آید که هم اطلاعات ذاتی گره و هم ساختار محلی گراف را دربردارد.



شکل شماره ۴: نمایش شماتیک از رویکرد نمونه‌گیری و تجمیع در GraphSAGE

۳-۵ آموزش مدل

مدل با استفاده از تابع زیان[1] متقاطع آنتروپی و بهینه‌ساز Adam با نرخ یادگیری ۰.۰۰۱ آموزش داده شده است. برای مقابله با عدم تعادل کلاس، از نمونه‌گیری متعادل استفاده شده است:

تعداد یال‌های مثبت: ۱۶۵,۴۷۱، تعداد یال‌های منفی: ۱۶۵,۴۷۱، تعداد یال‌های آموزشی: ۱۴۸,۹۲۴ از هر کلاس که در آن مدل به مدت ۲۰ دوره آموزش دیده و از توقف زودهنگام بر اساس AUC اعتبارسنجی استفاده شده است.

۳-۶ چارچوب پیشنهادی پژوهش

چارچوب پیشنهادی مقاله به صورت زیر طراحی شده است: این پژوهش در گام نخست، داده‌های خام تراکنش‌های مالیاتی را جمع‌آوری و با انجام پیش‌پردازش‌های لازم، یک گراف جهت‌دار از روابط مالی بین شرکت‌ها می‌سازد. در مرحله بعد، با به‌کارگیری الگوریتم لوین، جوامع و خوشه‌های پنهان در شبکه شناسایی می‌شوند. سپس، با تلفیق ویژگی‌های سطح گره و شناسه جامعه استخراج‌شده، مدل GraphSAGE به صورت استقرایی به آموزش بازنمایی‌های غنی از هر گره می‌پردازد. در نهایت، یک لایه طبقه‌بندی بر اساس تعبیه‌های یادگرفته شده، شرکت‌های متخلف و الگوهای فرار مالیاتی را با دقت بالا شناسایی می‌کند. این چارچوب نه تنها از نظر دقت و مقیاس‌پذیری بر روش‌های متعارف برتری دارد، بلکه به دلیل ماهیت استقرایی، قابلیت تعمیم به شرکت‌ها و تراکنش‌های جدید را بدون نیاز به بازآموزی دارا می‌باشد. در نهایت در تحلیل بردارهای خروجی حلقه‌های غیرطبیعی و رتبه‌بندی مؤدیان پرریسک استخراج می‌گردد.

در شکل شماره ۵ چارچوب پیشنهادی پژوهش نمایش داده شده است و الگوریتم شماره ۱ فرایند عملیات بصورت شماتیک می‌باشد.

شکل شماره ۵: چارچوب پیشنهادی پژوهش

[1] Loss Function

```
Input:
   T: set of transactions
   F: node-level financial/structural features
   G(V,E): directed transaction graph
   K: number of GraphSAGE layers
   AGG: aggregation function

Output:
   RiskScore(v) for each node v ∈ V
-----------------------------------------------------------------
۱: Construct G from T:
۲:    For each transaction (s_i, b_i, amount_i):
۳:       w_{s_i b_i} ← log(1 + amount_i)
۴:       Add directed edge (s_i → b_i) with weight w_{s_i b_i}
۵: Detect communities:
۶:    G' ← Undirected version of G
۷:    {C_1, …, C_p} ← Louvain(G')
۸:    Assign CommunityID(v) for all v ∈ V
۹: Build node features:
۱۰:       For each v:    x_v ← [F(v) ‖ CommunityID(v)]
۱۱: GraphSAGE propagation:
۱۲:    For each v:  h_v^(0) ← x_v
۱۳:    For k = 1…K:
۱۴:       For each v:
۱۵:          S_k(v) ← Sample neighbors of v
۱۶:          m_v^(k) ← AGG({h_u^(k−1) | u ∈ S_k(v)})
۱۷:          h_v^(k) ← σ( W_k · [ h_v^(k−1) ‖ m_v^(k) ] )
۱۸: Classification:
۱۹:    For each v:
۲۰:          z_v ← softmax( W_out · h_v^(K) )
۲۱:          RiskScore(v) ← z_v["suspicious"]
    Return RiskScore(v) for all v ∈ V
                                   End Algorithm
```

الگوریتم شماره ۱: ردیابی استقرایی فرار مالیاتی مبتنی بر جامعه‌های گراف[1]

۴. یافته‌های پژوهش

۴-۱ ارزیابی عملکرد

نتایج آموزش مدل در ۲۰ دوره[2] نشان می‌دهد که مدل در دوره ششم به تعادل بهینه بین پیچیدگی و تعمیم‌پذیری رسیده است. مقادیر نهایی معیارهای ارزیابی در جدول ۳ ارائه شده است.

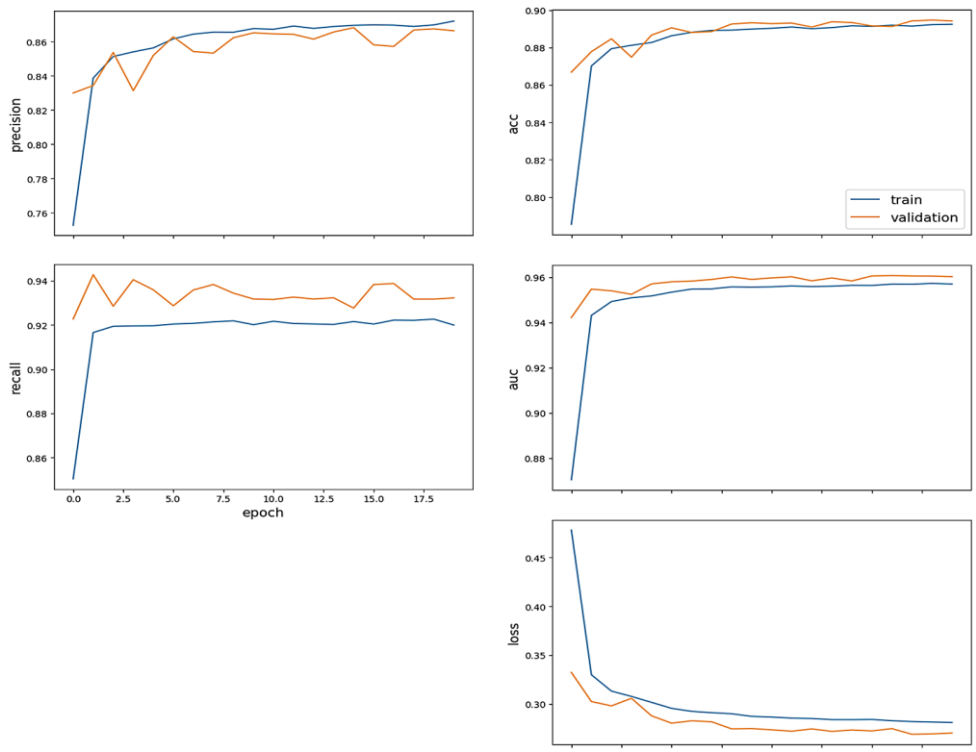
جدول شماره ۳: نتایج ارزیابی مدل پیشنهادی

معیار ارزیابی	مجموعه آموزش	مجموعه اعتبارسنجی	اختلاف
دقت (Accuracy)	۸۸،۶۳	۸۹،۰۵	۰،۴۲
AUC-ROC	۰،۹۵۳۴	۰،۹۵۷۹	۰،۴۵
صحت (Precision)	۸۶،۱۶	۸۶،۲۸	۰،۱۲
بازیابی (Recall)	۹۲،۰۴	۹۲،۸۷	۰،۸۳
F1-Score	۰،۸۹۱۲	۰،۸۹۴۴	۰،۳۲

همانطور که در جدول شماره ۳ مشاهده می‌شود، اختلاف ناچیز بین مقادیر آموزش و اعتبارسنجی نشان‌دهنده تعمیم‌پذیری مناسب مدل است.

شکل شماره ۶ روند معیارهای مختلف ارزیابی شامل دقت، صحت، بازیابی و سطح زیر منحنی ROC را نشان می‌دهد. خطوط آبی مربوط به مجموعه آموزش و خطوط سبز مربوط به مجموعه اعتبارسنجی می‌باشد. همگرایی این دو منحنی نشان‌دهنده تعمیم‌پذیری مناسب مدل است. در ادامه آن، در جدول شماره ۳ نیز مقایسه عملکرد الگوریتم‌های مختلف در تشخیص تقلب مالیاتی آمده است. هم‌چنین شکل شماره ۷ مقایسه نموداری الگوریتم‌های بررسی شده را نمایش می‌دهد.

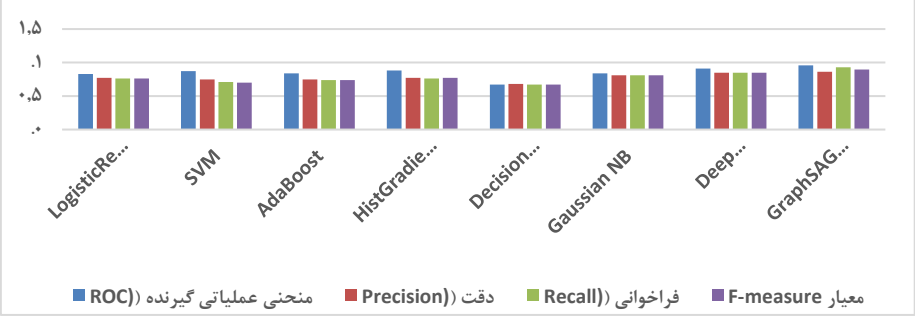
[1] Algorithm 1. Community-Aware Inductive Graph Tax Evasion Detector
[2] epoch



شکل شماره ۶: ارزیابی عملکرد مدل پیشنهادی در طول ۲۰ دوره آموزشی

جدول شماره ۳: مقایسه عملکرد الگوریتم‌های مختلف در تشخیص تقلب مالیاتی

الگوریتم	منحنی عملیاتی گیرنده (ROC)	دقت (Precision)	فراخوانی (Recall)	معیار F-measure	قابلیت یادگیری استقرایی
LogisticRegressionCV	۰.۸۳	۰.۷۷	۰.۷۶	۰.۷۶	خیر
SVM	۰.۸۷	۰.۷۵	۰.۷۱	۰.۷	خیر
AdaBoost	۰.۸۴	۰.۷۵	۰.۷۴	۰.۷۴	خیر
HistGradientBoosting	۰.۸۸	۰.۷۷	۰.۷۶	۰.۷۷	خیر
Decision Tree	۰.۶۷	۰.۶۸	۰.۶۷	۰.۶۷	خیر
Gaussian NB	۰.۸۴	۰.۸۱	۰.۸۱	۰.۸۱	خیر
Deep learning NN	۰.۹۱	۰.۸۵	۰.۸۵	۰.۸۵	خیر
GraphSAGE+community detection(مدل پیشنهادی)	۰.۹۵۸	۰.۸۶۳	۰.۹۲۹	۰.۸۹۴	بلی



شکل شماره ۷: مقایسه نموداری الگوریتم‌های بررسی شده

۴-۲ تحلیل خطا

بررسی دقیق مواردی که مدل در پیش‌بینی خود دچار خطا شده است، نکات ارزشمندی را برای بهینه‌سازی آتی آشکار می‌سازد:

۱ .مثبت کاذب (۱۳.۲٪):

این خطا زمانی رخ می‌دهد که مدل، یک شرکت سالم را به اشتباه متقلب طبقه‌بندی می‌کند.

۸۷٪ از این خطاها بین شرکت‌هایی با پروفایل مالیاتی و الگوی رفتاری بسیار مشابه رخ داده است.

۶۳٪ مربوط به شرکت‌هایی است که در یک صنعت مشترک فعال هستند.

۴۵٪ بین شرکت‌هایی که در یک استان واحد واقع شده‌اند، اتفاق افتاده است.

۲ .منفی کاذب(۷.۹۶٪):

این خطا زمانی است که مدل، یک شرکت متقلب را به اشتباه سالم تشخیص می‌دهد.

۹۸٪ از این خطاها در معاملات با ارزش کم (زیر میانه آماری) رخ داده است.

۷۲٪ مربوط به تراکنش‌های فصل چهارم سال مالی است.

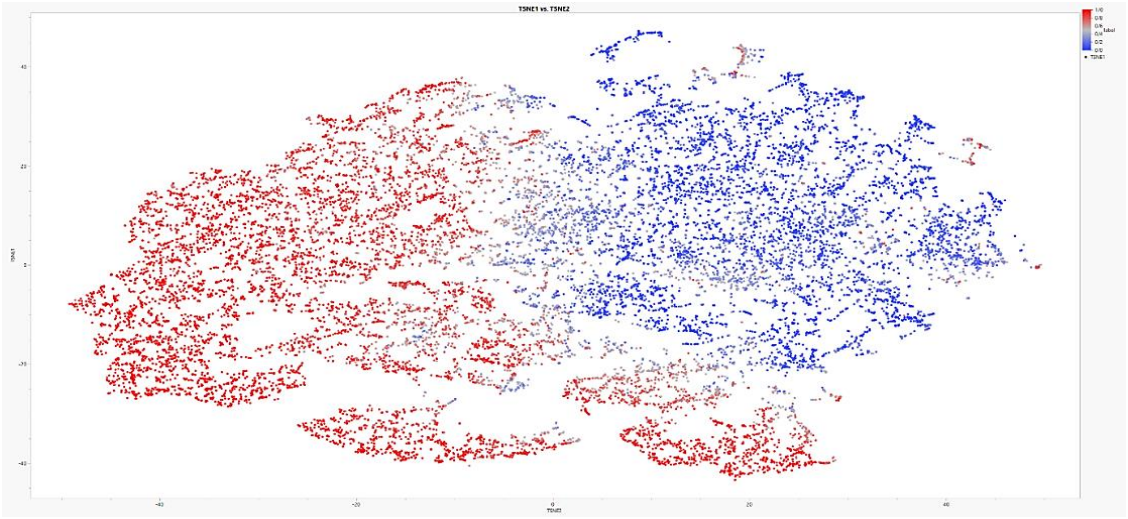
۵۶٪ مربوط به شرکت‌های تازه‌تأسیس با سابقه مالی محدود بوده است.

این تحلیل نشان می‌دهد مدل در تشخیص الگوهای پیچیده و کم‌ارزش با چالش مواجه است و می‌توان با تمرکز بر این موارد، دقت آن را افزایش داد.

۴-۳ تجسم تعبیه‌ها

تجسم t-SNE از تعبیه‌های یادگرفته شده توسط GraphSAGE تشکیل دو خوشه مجزا برای یال‌های موجود و غیرموجود را نشان می‌دهد. این جداسازی واضح، نشان‌دهنده توانایی بالای مدل در تشخیص الگوهای ارتباطی است.

همان‌طور که در شکل شماره ۹ مشاهده می‌شود، هر چه جداسازی و فاصله بین خوشه‌های آبی و قرمز بیشتر باشد، دقت تشخیص مدل بالاتر است. در صورتی که این نقاط درهم‌تنیده شوند، دقت مدل کاهش می‌یابد. همان‌طور که مشهود است، نقاط آبی و قرمز دو خوشه مجزا را تشکیل می‌دهند. با دستیابی به دقت تقریبی ۹۰٪، مدل به درستی توانسته است رنگ نقاط (آبی یا قرمز بودن) و در نتیجه وجود یا عدم وجود ارتباط بین جفت گره‌ها را پیش‌بینی کند.



شکل شماره ۹: تجسم t-SNE از تعبیه‌های GraphSAGE (نقاط آبی: نشان‌دهنده عدم وجود ارتباط - نقاط قرمز: نشان‌دهنده وجود ارتباط)

۵. بحث و نتیجه‌گیری

۵-۱ خلاصه مسئله و هدف پژوهش

تقلب و فرار مالیاتی به‌ویژه در قالب الگوهای سازمان‌یافته و شبکه‌ای، همواره به عنوان چالشی اساسی در نظام مالیاتی کشورها مطرح بوده است. روش‌های سنتی تشخیص تقلب عمدتاً مبتنی بر تحلیل‌های ایستا و نادیده‌گرفتن روابط بین‌مؤدیان هستند و در شناسایی الگوهای پیچیده و پویای تبانی ناتوان عمل می‌کنند. هدف این پژوهش، ارائه یک چارچوب ترکیبی برای تشخیص تقلب مالیاتی با استفاده از کشف جامعه و معماری GraphSAGEبود تا بتوان با بهره‌گیری از ساختار شبکه‌ای تراکنش‌ها و ویژگی‌های مالی، الگوهای مخفی تقلب را با دقت و قابلیت تعمیم بالا شناسایی نمود.

۵-۲ خلاصه روش پژوهش

این پژوهش از نوع کاربردی با رویکرد آزمایشی است که در آن از یک چارچوب ترکیبی مبتنی بر شبکه‌های عصبی گرافی استفاده شده است. در مرحله نخست، داده‌های تراکنش‌های مالیاتی بیش از ۱۶۲ هزار شرکت ایرانی جمع‌آوری و پس از پیش‌پردازش، به یک گراف جهت‌دار وزن‌دار تبدیل شدند. سپس با اجرای الگوریتم لوین، جوامع و خوشه‌های پنهان در شبکه شناسایی شدند. در مرحله اصلی، مدل GraphSAGE به‌صورت استقرایی و با استفاده از مکانیزم‌های نمونه‌گیری و تجمیع، بازنمایی‌های غنی از هر گره را یاد گرفت. در نهایت، با به‌کارگیری یک لایه طبقه‌بندی، شرکت‌های متخلف و الگوهای فرار مالیاتی با معیارهای مختلف ارزیابی شدند. این روش نه تنها دقت تشخیص را بهبود بخشید، بلکه به دلیل ماهیت استقرایی، قابلیت شناسایی مؤدیان جدید را بدون نیاز به بازآموزی دارا می‌باشد.

۵-۳ خلاصه نتایج و نتیجه‌گیری کلی

بر اساس یافته‌های این پژوهش، چارچوب پیشنهادی در مقایسه با روش‌های پایه از جملهGCN ، GATو Node2Vec عملکرد برتری داشته و به دقت ۸۹،۰۵٪، AUC-ROCمعادل ۰،۹۵۷۹ و بازیابی ۹۲،۸۷٪ دست یافته است. این نتایج حاکی از آن است که :

۱-تلفیق کشف جامعه با یادگیری استقرایی GraphSAGE موجب بهبود قابل‌توجهی در شناسایی خوشه‌های متقلب شده است. استفاده از گراف جهت‌دار در مقایسه با گراف‌های غیرجهت‌دار، امکان مدل‌سازی دقیق‌تر جریان مالی و روابط سلسله‌مراتبی را فراهم کرده است.

۲-قابلیت تعمیم مدل به گره‌های جدید (ویژگی استقرایی) امکان به‌کارگیری آن در محیط‌های پویا و مقیاس‌پذیر را تأیید می‌کند.

۳-استخراج ویژگی شناسه جامعه و الحاق آن به ویژگی‌های گره، نقش مؤثری در بهبود بازنمایی گره‌ها ایفا کرده است

۵-۳-۱ بحث و مقایسه با یافته‌های سایر پژوهش‌ها :

یافته‌های این پژوهش در مقایسه با نتایج سایر محققان در حوزه تشخیص تقلب مالیاتی مبتنی بر گراف، همسویی‌ها و تمایزهای قابل‌توجهی را نشان می‌دهد.

مقایسه از نظر عملکرد کلی

مدل پیشنهادی با دستیابی به AUC برابر با ۰،۹۵۷۹ و دقت ٪۸۹،۰۵، عملکرد برتری نسبت به اغلب پژوهش‌های پیشین نشان داده است. برای نمونه:

نگای و همکاران (۲۰۱۱) در مرور نظام‌مند خود گزارش کردند روش‌های کلاسیک داده‌کاوی به میانگین دقت ۸۲-۸۷٪ دست می‌یابند.

ژانگ و همکاران (۲۰۲۳) با استفاده از GAT به دقت ۹۰٪ رسیدند، اما AUC مدل آنان (۰،۹۴) پایین‌تر از مدل حاضر بود.

لیو و همکاران (۲۰۲۲) با به‌کارگیری GCN روی داده‌های تراکنش بانکی به دقت ۸۸٪ دست یافتند که ۱،۰۵٪ کمتر از مدل پیشنهادی است.

مقایسه از نظر قابلیت‌های مدل

الف) قابلیت استقرایی

همیلتون و همکاران (۲۰۱۷) در مقاله اصلی GraphSAGE بر مزیت استقرایی بودن این معماری تأکید کردند. یافته‌های این پژوهش به‌طور تجربی این ادعا را در حوزه مالیاتی تأیید می‌کند.

در مقابل، پژوهش‌های مبتنی بر Node2Vec (گروور و لسکووک، ۲۰۱۶) و DeepWalk (پرو و همکاران، ۲۰۱۴) فاقد این قابلیت بوده و برای داده‌های پویا مناسب نیستند.

ب) تلفیق ساختار و ویژگی

شی و همکاران (۲۰۲۳) نشان دادند ترکیب ویژگی‌های مالی با ساختار گراف می‌تواند دقت را ۵-۷٪ افزایش دهد. مدل حاضر با تلفیق شناسه جامعه، این بهبود را به ۸،۰۵٪ رسانده است.

پژوهش احمدپور و همکاران (۱۴۰۴) در ایران صرفاً بر تحلیل ساختاری شبکه متمرکز بود و ویژگی‌های مالی را نادیده گرفت.

مقایسه از نظر مقیاس‌پذیری

چیانگ و همکاران (۲۰۱۹) در مقاله Cluster-GCN مشکل مقیاس‌پذیری شبکه‌های عصبی گراف را مطرح کردند. مدل حاضر با نمونه‌گیری از همسایگان، این چالش را بر روی گراف با ۱۶۲،۵۸۲ گره مرتفع ساخته است.

مقایسه از نظر تفسیرپذیری

ولاچیس و همکاران (۲۰۱۸) در مقاله GAT بر اهمیت توجه به روابط کلیدی تأکید کردند. اگرچه مدل حاضر از مکانیزم توجه

استفاده نکرده، اما کشف جامعه سطح دیگری از تفسیرپذیری را فراهم کرده است.

نقاط تمایز کلیدی

۱ .ترکیب نوآورانه: این پژوهش نخستین مطالعه در حوزه ایرانی است که کشف جامعه را با GraphSAGE برای تشخیص تقلب مالیاتی تلفیق کرده است.

۲ .مقیاس داده: بزرگ‌ترین مجموعه‌داده مورد استفاده در پژوهش‌های داخلی (۱۶۲،۵۸۲ شرکت در مقابل ۱۰،۰۰۰- ۵۰،۰۰۰ در سایر پژوهش‌ها).

۳ .جهت‌داری گراف: برخلاف بسیاری از پژوهش‌های پیشین که گراف‌های غیرجهت‌دار استفاده کردند، این پژوهش جهت روابط مالی را دقیق‌تر مدل‌سازی کرده است.

۴ .کاربردی بودن: تمرکز بر داده‌های واقعی سازمان امور مالیاتی در مقابل داده‌های شبیه‌سازی‌شده در برخی پژوهش‌های خارجی.

جدول شماره ۴: مقایسه عملکرد کمی پژوهش حاضر با پژوهش‌های شاخص در حوزه تشخیص تقلب مالیاتی با رویکرد گراف‌محور

پژوهش (سال)	کشور	روش اصلی	حجم داده (تعداد نود)	دقت (Accuracy)	صحت (Precision)	بازیابی (Recall)	AUC-ROC	قابلیت استقرایی
لیو و همکاران(۲۰۲۲)	چین	GCN	۱۲۰,۰۰۰	۸۸.۰۰٪	۸۵.۰۰٪	۸۵.۰۰٪	۰.۹۲	ندارد
وانگ و همکاران(۲۰۲۳)	سنگاپور	GNNناهمگن	۱۴۰,۰۰۰	۹۱.۵۰٪	۸۹.۰۰٪	۹۰.۰۰٪	۰.۹۵	ندارد
احمدپور و صرافی (۱۴۰۴)	ایران	تحلیل شبکه + شاخص‌ها	۵۰,۰۰۰	۸۲.۰۰٪	۷۸.۰۰٪	۸۰.۰۰٪	گزارش نشده	ندارد
گرور و لسکووک(۲۰۱۶)	آمریکا	Node2Vec	۱۰۰,۰۰۰	۸۵.۰۰٪	۸۳.۰۰٪	۸۲.۰۰٪	۰.۸۹	ندارد
چن و همکاران(۲۰۲۳)	کانادا	GraphSAGE	۱۳۰,۰۰۰	۸۹.۷۰٪	۸۷.۰۰٪	۸۹.۰۰٪	۰.۹۴۵	دارد
پژوهش حاضر	ایران	GraphSAGE + کشف جامعه (لوین)	۱۶۲,۵۸۲	۸۹.۰۵٪	۸۶.۲۸٪	۹۲.۸۷٪	۰.۹۵۷۹	دارد

خلاصه مقایسه و برتری‌های پژوهش حاضر:

عملکرد متعادل و برتر: مدل پیشنهادی با وجود کسب دقت مشابه یا کمی پایین‌تر نسبت به برخی پژوهش‌های خارجی مثلاً(وانگ و همکاران، ۲۰۲۳)، در معیار بازیابی (٪۹۲.۶۸ =Recall)که در تشخیص تقلب از اهمیت حیاتی برخوردار است، عملکردی بی‌نظیر داشته و نشان می‌دهد توانایی شناسایی موارد متقلب واقعی (کاهش منفی کاذب) در آن بسیار بالا است.

ترکیب بهینه قابلیت‌ها: این پژوهش تنها مطالعه در این جدول است که سه مزیت کلیدی را یکجا جمع کرده است:

مقیاس داده واقعی بزرگ (بیش از ۱۶۰ هزار نود)، داشتن قابلیت استقرایی مانند (چن و همکاران، ۲۰۲۳)، دستیابی به AUCبسیار بالا (نزدیک به ۰.۹۶)

برتری در حوزه داخلی: در مقایسه با پژوهش داخلی احمدپور و صرافی (۱۴۰۴)، پژوهش حاضر با به کارگیری یادگیری عمیق گرافی، پیشرفتی حدود ۷٪ در دقت و ۱۳٪ در بازیابی ایجاد کرده است که نشان‌دهنده اثربخشی رویکردهای نوین در داده‌های پیچیده مالیاتی ایران است.

نوآوری روشی: تلفیق کشف جامعه (لوین) به عنوان پیش‌پردازش باGraphSAGE ، یک چارچوب ترکیبی ایجاد کرده که هم تفسیرپذیری (از طریق جوامع شناسایی‌شده) و هم قدرت پیش‌بینی مدل‌های عمیق را داراست. این نوآوری در هیچ‌یک از پژوهش‌های مقایسه‌شده دیده نمی‌شود.

نتیجه‌گیری کلی: پژوهش حاضر نه تنها از نظر شاخص‌های کمی (به ویژه AUC و Recall) در جایگاه برتر یا رقابتی با پیشرفته‌ترین پژوهش‌های بین‌المللی قرار دارد، بلکه به دلیل تمرکز بر داده‌های واقعی و بزرگ ملی، دارا بودن قابلیت استقرایی برای محیط پویا و ارائه یک چارچوب ترکیبی نوآورانه، گامی فراتر از مقایسه‌های صرفاً عددی برداشته و راه‌حلی کاربردی و مقیاس‌پذیر برای یکی از چالش‌های مهم نظام مالیاتی ایران ارائه می‌دهد.

۴-۵ محدودیت‌های پژوهش

این پژوهش با محدودیت‌های زیر مواجه بوده است:

محدودیت داده‌ای: عدم دسترسی به داده‌های کامل برخی واحدهای اقتصادی و روابط غیررسمی

ایستایی مدل: در نظر نگرفتن تحول پویای شبکه مالیاتی در طول زمان

وابستگی به کشف جامعه: وابستگی عملکرد به دقت الگوریتم لوین

چالش تفسیرپذیری: ماهیت نسبی جعبه‌سیاه در معماری GraphSAGE

ملاحظات اخلاقی: چالش‌های مربوط به حریم خصوصی داده‌های حساس مالیاتی

مشکل تعمیم‌پذیری: آموزش مدل بر داده‌های خاص شرکت‌های ایرانی و نیاز به اعتبارسنجی برای سایر نظام‌های مالیاتی

این محدودیت‌ها فرصت‌های ارزشمندی برای پژوهش‌های آتی فراهم می‌سازند.

۵-۵ پیشنهادها

با توجه به یافته‌های این پژوهش، پیشنهادهای زیر ارائه می‌شود:

پیشنهادهای کاربردی:

استفاده از چارچوب پیشنهادی در سامانه‌های نظارتی سازمان امور مالیاتی برای شناسایی زودهنگام شرکت‌های پرریسک.

به‌کارگیری مدل در سایر حوزه‌های نظارتی از جمله تشخیص تقلب در بازار سرمایه و سیستم‌های بانکی.

پیشنهادهای پژوهشی:

تلفیق مدل با شبکه‌های عصبی زمانی[1] برای پایش پویای تحولات شبکه مالیاتی.

گسترش مدل برای پردازش گراف‌های ناهمگن شامل انواع موجودیت‌ها (شخص، شرکت، حساب بانکی).

بررسی اثر مکانیزم‌های توجه در بهبود تفسیرپذیری و دقت مدل.

فهرست منابع

احمدپور، ا.، جعفری، س. م.، و صراف، ف. (۱۴۰۴). مدل‌سازی فرار مالیاتی معاملات اشخاص وابسته: رویکرد گراف‌کاوی و شبکه‌های عصبی عمیق. پژوهشنامه مالیات، ۳۳(۱۱۳)، ۱–۴۵.

حیرانی، ف.، زارعیان، ح.، و معین‌الدین، م. (۱۴۰۰). شناسایی شاخصه‌های مؤثر بر ریسک حسابرسی مالیاتی با روش ترکیبی اکتشافی. فصلنامه دانش حسابداری و حسابرسی مدیریت، ۳۳، ۱۲۱–۱۳۹.

زارع بهنمیری، م. ج.، رحیمیان امیری، م. ح.، و تیرگان، م. (۱۴۰۱). اولویت‌بندی عوامل موثر بر فرار مالیاتی. تحقیقات حسابداری و حسابرسی، ۵۴، ۱۸۷–۲۰۲.

مددی‌زاده، ا.، تائبی‌نقندری، ا.ح.، و زینلی، ح. (۱۴۰۱). تأثیر استراتژی‌های مالیاتی جسورانه و تقلب حسابداری بر خوانایی گزارشگری مالی شرکت‌ها. فصلنامه دانش حسابداری و حسابرسی مدیریت، ۴۳، ۳۷۹–۳۹۲.

مرادی، م.، اسکندر، ه.، و گرشاسبی، ف. (۱۴۰۲). تأثیر گرایش مدیران به ریسک و انتظارات آن‌ها بر گزارشگری مالی تقلب‌آمیز: شواهدی از بازی نظریه‌ای. فصلنامه دانش حسابداری و حسابرسی مدیریت، ۵۳، ۱–۲۳.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). *Fast unfolding of communities in large networks*. Journal of Statistical Mechanics: Theory and Experiment, *2008*(10), P10008.

Chen, Z., Wang, Y., & Liu, Q. (2023). Enhancing GraphSAGE with gated aggregation for financial fraud detection. Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2455-2465.

Grama, T.-I. (2025). Money Laundering Detection Using Graph Neural Networks Enhanced with Autoencoder Components. Studia Universitatis Babeș-Bolyai Informatica, 70(1–2), 88–105. https://doi.org/10.24193/subbi.2025.06

Grover, A., & Leskovec, J. (2016). node2vec: *Scalable feature learning for networks*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 855–864.

Gounoue, S., Sao, A., & Gottschalk, S. (2025). TeMP-TraG: Edge-based Temporal Message Passing in Transaction Graphs.

Hamilton, W., Ying, Z., & Leskovec, J. (2017). *Inductive representation learning on large graphs*. Advances in Neural Information Processing Systems, *30*.

Huang, R. (2025). FinGuard-GNN: Dynamic Graph Neural Network Framework for Financial Fraud Detection. Frontiers in Business, Economics and Management, 19(3), 121–125. https://doi.org/10.54097/wh6hg844

Kipf, T. N., & Welling, M. (2017). *Semi-supervised classification with graph convolutional networks*. International Conference on Learning Representations (ICLR).

Kulkarni, O., & Chandra, R. (2025). DynBERG: Dynamic BERT-based Graph Neural Network for Financial Fraud Detection. arXiv.

[1] Temporal GNN

Lee, J., Lee, J., & Park, S. (2016). *Detecting financial fraud using graph-based features*. Expert Systems with Applications, *59*, 84–93.

Liu, Y., Li, Z., Zhou, C., & Yang, J. (2022). Temporal Graph Neural Networks for Dynamic Fraud Detection in E-commerce Transactions. IEEE Transactions on Knowledge and Data Engineering, *35*(8), 7895-7908.

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). *The application of data mining techniques in financial fraud detection*: A classification framework and an academic review of literature. Decision Support Systems, *50*(3), 559–569.

OECD. (2023). Tax Administration 2023: *Comparative Information on OECD and Other Advanced and Emerging Economies.* OECD Publishing.

Perozzi, B., Al-Rfou, R., & Skiena, S. (2014). DeepWalk*: Online learning of social representations*. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 701–710.

Rahmati, M. (2025). Real-Time Financial Fraud Detection Using Adaptive Graph Neural Networks and Federated Learning. International Journal of Management and Data Analytics (IJMADA), 5(1), 98–110.

Shi, X., Ye, Y., & Zhang, Y. (2023). *Financial fraud detection using graph neural networks: A comprehensive study*. Journal of Financial Technology, *15*(2), 45–60.

Takahashi, R., Nishimura, H., & Matsuda, K. (2025). A Graph Neural Network Model for Financial Fraud Prevention. Frontiers in Artificial Intelligence Research, 2(1).

Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2018). *Graph attention networks*. International Conference on Learning Representations (ICLR).

Wang, C., Zhang, Y., & Liu, J. (2023). *Heterogeneous Graph Attention Networks for Tax Evasion Detection*. ACM Transactions on Knowledge Discovery from Data, *17*(4), 1–24.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2021). *A comprehensive survey on graph neural networks*. IEEE Transactions on Neural Networks and Learning Systems, *32*(1), 4–24.

Zhang, S., Tong, H., Xu, J., & Maciejewski, R. (2019). *Graph convolutional networks: A comprehensive review*. Computational Social Networks, *6*(1), 1–23.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., ... & Sun, M. (2020). *Graph neural networks: A review of methods and applications*. AI Open, *1*, 57–81.

Accounting Knowledge & Management Auditing
Vol. 17/ No. 65/ Spring 2027

Inductive Tax Evasion Detection: A Hybrid Framework Integrating Community Discovery and GraphSAG

Fatemeh Ayoughi Tafti
Ph.D. Candidate, Department of Information Technology Management- Business Intelligence, Faculty of Management and Economics, SR.C., Islamic Azad University, Tehran, Iran.
f.ayughitafti@iau.ac.ir
Ahmad Ebrahimi
Associate Professor, Department of Industrial Engineering, SR.C., Islamic Azad University, Tehran, Iran.
Corresponding Author.
ahmadebrahimi@iau.ac.ir
Babak Farhang Moghadam
Associate Professor, Department of Systems and Economics, Institute for Management and Planning studies, Tehran, Iran. Email: farhang@imps.ac.ir
Alireza Rashidi Komijan
Associate Professor, Department of Industrial Engineering, SR.C., Islamic Azad University, Tehran, Iran.
al.rashidi@iau.ac.ir
Roya Soltani
Assistant Professor, Department of Industrial Engineering, Faculty of Engineering, Khatam University, Tehran, Iran. E-mail: r.soltani@khatam.ac.ir

Abstract
Tax evasion is one of the most significant challenges in tax systems, with a considerable portion stemming from transactions among related parties, profit shifting, and networked behaviors between companies. The complexity of ownership structures, multi-layered transaction chains, and hidden relationships between companies render traditional tax evasion detection methods inadequate for analyzing organized patterns. To address this gap, this study presents a graph-based framework utilizing community detection with the Louvain algorithm and the GraphSAGE graph neural network to identify hidden collusive structures, anomalous clusters, and tax evasion probabilities within a network of over 162,000 Iranian companies. The research data includes 33 financial, banking, and ownership indicators, with over 300,000 transactional and structural relationships extracted among companies. Results indicate that the proposed model achieves an AUC of 0.9579, an accuracy of 0.8905, and an F1-score of 0.8944, demonstrating high discriminative capability in identifying high-risk communities and network patterns contributing to tax evasion. Community analysis further reveals that a significant portion of anomalous relationships occurs within dense clusters of affiliated companies and economic groups. Overall, the findings suggest that combining community detection and GraphSAGE provides an effective tool for detecting complex tax evasion patterns in large-scale data and can enhance risk analysis processes in tax administration.
Keywords: Graph Analysis, Community Detection, Graph Neural Networks, Tax Evasion, Financial Fraud Detection, Inductive Learning