ICIRES 2026
www.icires.ir
info@icires.ir
بیست و سومین کنفرانس بین المللی نوآوری و تحقیق در علوم مهندسی
گرجستان – اسفند ماه ۱۴۰۴
13 March 2026 -TBILISI GEORGIA

یک روش یادگیری بدون نظارت برای تشخیص الگوهای غیرمعمول و ناهنجاری در داده‌های تراکنشی

سارا نجفی
دانشجوی کارشناسی ارشد هوش مصنوعی دانشگاه آزاد واحد الکترونیکی
sara.njf.ch@gmail.com

دکتر مریم تعجبیان
استادیار مهندسی کامپیوتر دانشگاه آزاد واحد مهدی‌شهر
taajobian@msh-iau.ac.ir

دکتر سابینا نوبری
مدرس دانشگاه آزاد واحد الکترونیکی
Sabinamirzaei@gmail.com

چکیده

در عصر داده‌محور کنونی، تشخیص ناهنجاری در داده‌های تراکنشی نقش حیاتی در شناسایی رفتارهای غیرعادی، کشف تقلب، و ارتقاء امنیت سامانه‌های مالی و تجاری ایفا می‌کند. با توجه به ماهیت پیچیده و پویا بودن تراکنش‌ها، استفاده از روش‌های یادگیری بدون نظارت به‌عنوان رویکردی مؤثر برای کشف الگوهای غیرمعمول بدون نیاز به برچسب‌گذاری دستی، اهمیت فزاینده‌ای یافته است. در این پژوهش، یک چارچوب نوین مبتنی بر یادگیری بدون نظارت برای شناسایی ناهنجاری‌ها در داده‌های تراکنشی ارائه شده است. الگوریتم‌های متنوعی از جمله جنگل جداسازی(Isolation Forest) ، ماشین بردار پشتیبان یک‌کلاسه (One-Class SVM)و عامل دورافتاده محلی (Local Outlier Factor) به‌کار گرفته شده‌اند تا رفتارهای غیرعادی در داده‌های واقعی شناسایی شوند. داده‌ها پس از پیش‌پردازش و نرمال‌سازی، با استفاده از معیارهای ارزیابی مانند نرخ کشف ناهنجاری، دقت وF ۱-Scoreمورد تحلیل قرار گرفته‌اند. نتایج تجربی نشان می‌دهد که روش ترکیبی، توان بالایی (دقت ۹۹.۶٪) در شناسایی الگوهای ناهنجار در داده‌های تراکنشی دارد. این چارچوب می‌تواند به‌عنوان پایه‌ای برای توسعه سیستم‌های هوشمند کشف تقلب و تحلیل رفتاری در حوزه‌های مالی، بانکی و تجارت الکترونیک مورد استفاده قرار گیرد.

کلمات کلیدی:

واژگان کلیدی: روش یادگیری بدون نظارت، تشخیص الگوهای غیرمعمول، ناهنجاری، داده های تراکنشی، جنگل جداسازی.

مقدمه

کشف ناهنجاری‌ها (Anomaly Detection) در داده‌های تراکنشی یکی از موضوعات مهم و حیاتی در حوزه‌های مالی، امنیتی و تجاری است. این فرآیند به شناسایی رفتارها یا الگوهایی می‌پردازد که به طور قابل توجهی با رفتار عادی سیستم یا کاربران تفاوت دارند. اهمیت و ضرورت کشف ناهنجاری‌ها در داده‌های تراکنشی را می‌توان از جنبه‌های مختلفی بررسی کرد :جلوگیری از تقلب و کلاهبرداری مالی، افزایش امنیت سیستم‌ها و داده‌ها، بهبود تجربه کاربری، بهینه‌سازی فرآیندها و کاهش هزینه‌ها، انطباق با قوانین و مقررات، شناسایی فرصت‌های جدید، مقابله با فعالیت‌های غیرقانونی، افزایش دقت و کارایی سیستم‌ها(Bajpai, A. D. (2024).)
با توجه به توضیحات بالا، می توان کشف ناهنجاری‌ها در داده‌های تراکنشی از جنبه‌های مختلفی حائز اهمیت است، از جمله جلوگیری از تقلب، افزایش امنیت، بهبود تجربه کاربری، بهینه‌سازی فرآیندها و انطباق با قوانین. این فرآیند نه تنها به کاهش هزینه‌ها و خسارات مالی کمک می‌کند، بلکه می‌تواند به شناسایی فرصت‌های جدید و بهبود استراتژی‌های تجاری نیز منجر شود. با توجه به افزایش حجم و پیچیدگی داده‌های تراکنشی، استفاده از روش‌های پیشرفته یادگیری ماشین و تحلیل داده‌ها برای تشخیص ناهنجاری‌ها ضروری است. لذا در این تحقیق یک روش نوین با استفاده از ترکیب پنج الگوریتم یادگیری ماشین برای شناسایی سریعتر ناهنجاری در داده های تراکنشی پیشنهاد خواهیم کرد.
این مقاله در پنج بخش تنظیم شده است. در بخش دوم مروری بر تحقیقات انجام شده توسط محققان دیگر خواهیم داشت. در بخش سوم روش پیشنهادی را توضیح خواهیم داد. در بخش چهارم نتایج بدست آمده از پیاده سازی روش پیشنهادی و ارزیابی کارایی روش پیشنهادی را خواهیم داشت و در نهایت نتیجه گیری و محدودیت های تحقیق را خواهیم آورد.

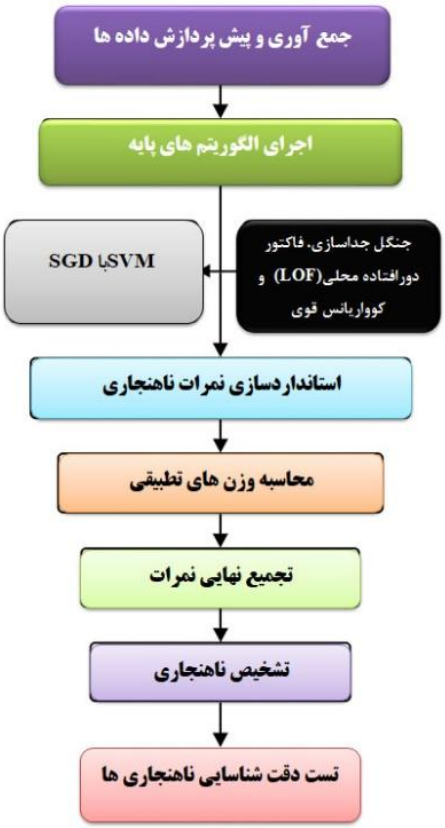
پیشینه تحقیق

Bolton و همکارانش.((Bolton, R. J. (2021) به کشف تقلب رفتاری از طریق تجزیه و تحلیل داده های طولی پرداختند. این داده‌ها معمولاً از تراکنش‌های کارت اعتباری در طول زمان تشکیل می‌شوند، اما می‌توانند متغیرهای دیگری، هم ثابت و هم طولی را شامل شوند. آنها در این مقاله دو روش (آماری و روشهای بدون نظارت) را برای کشف تقلب بدون نظارت در داده‌های اعتباری مورد بحث قرار دادند و آن‌ها را برای برخی از مجموعه‌های داده واقعی اعمال کردند. برای پرداختن به ریسک نمونه‌گیری و ناکارآمدی حسابرسی مالی، Bakumenko و همکارانش (.Bakumenko, A. a. (2022)) از هفت تکنیک ML نظارت شده شامل یادگیری عمیق و دو تکنیک ML بدون نظارت مانند جنگل جداسازی و رمزگذار خودکار استفاده کردند. آنها مدل‌های خود را بر روی یک مجموعه داده واقعی GL آموزش و ارزیابی کردند و از بردارسازی داده‌ها برای حل تنوع اندازه ورودی مجله استفاده کردند. نتایج ارزیابی نشان داد که بهترین مدل‌های تحت نظارت و بدون نظارت آموزش‌دیده، پتانسیل بالایی در تشخیص انواع ناهنجاری از پیش تعریف‌شده و همچنین در نمونه‌برداری کارآمد از داده‌ها برای تشخیص ورودی‌های مجله با ریسک بالاتر دارند. Gogoi و همکارانش (Gogoi, P. B. (2020)) بر روی ANIDS های موجود را بر اساس نوع، کلاس، ماهیت شناسایی/پردازش، سطح امنیت، و غیره دسته بندی کردند. همچنین برخی از اقدامات نزدیکی را برای تجزیه و تحلیل و تشخیص داده های نفوذ به کار گرفتند. آنها همچنین برخی از نتایج تجربی را برای شناسایی حملات بر روی مجموعه داده '۹۹KDD گزارش نمودند. Schlör و همکارانش ((Schlör, D. (2022)
تکنیک‌های مختلف یادگیری نمایش را برای تبدیل داده‌های تراکنش طبقه‌ای به بردارهای عددی متراکم بررسی کردند. آنها این رویکرد را با پیشنهاد گسسته‌سازی آگاهانه دورتر گسترش دادند. در مرحله بعد، سناریوهای مختلف را برای تشخیص ناهنجاری در داده‌های تراکنش ارزیابی نمودند. آنها معماری iNALU خود را به یک لایه عصبی گسترش دادند که می‌تواند وابستگی‌های عددی و غیرعددی را مدل‌سازی کند و آن را در یک محیط نظارت شده و یک کلاسه ارزیابی کند. در ادامه ثبات و قابلیت تعمیم رویکرد خود را بررسی کردند و نشان دادند که عملکرد آن از انواع مدل‌ها در تنظیمات نظارت‌شده متعادل و عملکرد قابل‌مقایسه‌ای در

تنظیمات یک کلاسه بهتر است. در نهایت، آنها سه رویکرد برای استفاده از یک مدل تولیدی به عنوان آشکارساز ناهنجاری را ارزیابی نمودند و عملکرد تشخیص ناهنجاری را مقایسه کردند. Munir و همکارانش (Munir, M. e. (2018).) یک رویکرد جدید تشخیص ناهنجاری مبتنی بر یادگیری عمیق (DeepAnT) برای داده‌های سری زمانی ارائه کردند که به همان اندازه برای موارد غیر جریانی قابل اجرا است. DeepAnT قادر به تشخیص طیف گسترده ای از ناهنجاری ها، به عنوان مثال، ناهنجاری های نقطه، ناهنجاری های متنی، و اختلاف در داده های سری زمانی است. برخلاف روش‌های تشخیص ناهنجاری که در آن ناهنجاری‌ها آموخته می‌شوند، DeepAnT از داده‌های بدون برچسب برای ضبط و یادگیری توزیع داده استفاده می‌کند که برای پیش‌بینی رفتار عادی یک سری زمانی استفاده می‌شود. Cholevas و همکارانش ((Cholevas, C. e. (2024) تکنیک‌های تشخیص ناهنجاری در اکوسیستم‌های بلاک چین را از طریق دریچه یادگیری بدون نظارت، کاوش در پیچیدگی‌ها و گذر از مجموعه پیچیده رفتارهای غیرعادی با بررسی الگوریتم‌های آوانگارد برای تشخیص انحراف از الگوهای عادی بررسی کردند. آنها پیشنهاد کردند که استفاده از الگوریتم‌های نظارت‌نشده در تشخیص ناهنجاری بلاک‌چین نه تنها باید به‌عنوان یک رویه پیاده‌سازی، بلکه به‌عنوان یک رویه یکپارچه‌سازی در نظر گرفته شود، به طوری که مزایای این الگوریتم‌ها را می‌توان به‌طور مؤثری به روش‌هایی که با مشکل در دست تعیین می‌شود ترکیب کرد. این مقاله همچنین ارائه عمیقی از ساختارهای داده ای را ارائه می دهد که معمولاً در تشخیص ناهنجاری بلاک چین مبتنی بر یادگیری بدون نظارت استفاده می شود. Ball و همکارانش ((Ball, R. S. (2023.) یک چارچوب مبتنی بر یادگیری ماشین برای تشخیص ناهنجاری پیشنهاد کردند که هدف آن ترکیب چندین رویکرد یادگیری ماشین برای تشخیص ناهنجاری‌ها در سیستم‌های تراکنش است. چارچوب پیشنهادی شامل یک رویکرد واحد برای ترکیب جنبه‌های فرآیند تشخیص ناهنجاری در یک خط لوله منفرد است. این رویکرد با جستجوی معماری عصبی آغاز می‌شود، که در آن چندین معماری رمزگذار خودکار کاندید به‌طور تصادفی شبیه‌سازی می‌شوند تا یک پیکربندی معماری بهینه را تعیین کنند. چارچوب پیشنهادی نهایی در یک محیط واقعی برای نشان دادن کاربرد عملی اجرا شد. نتایج به‌دست‌آمده از پیاده‌سازی با موفقیت به هدف اصلی، که بهبود تشخیص ناهنجاری‌ها در یک محیط معاملاتی بود، دست یافت. Huang و همکارانش (Huang, D. e. (2018).)، یک چارچوب جدید کشف تقلب به نام CoDetect را پیشنهاد کردند که می‌تواند هم اطلاعات شبکه و هم اطلاعات ویژگی را برای تشخیص تقلب مالی اعمال کند. علاوه بر این، CoDetect می تواند به طور همزمان فعالیت های کلاهبرداری مالی و الگوهای ویژگی های مرتبط با فعالیت های کلاهبرداری را شناسایی کند. Carcillo و همکارانش(Carcillo, F. e. (2021).) یک تکنیک ترکیبی را ارائه کردند که تکنیک‌های نظارت شده و بدون نظارت را برای بهبود دقت تشخیص تقلب ترکیب می‌کند. امتیازات پرت بدون نظارت، محاسبه شده در سطوح مختلف جزئیات، با یک مجموعه داده کشف تقلب کارت اعتباری واقعی، حاشیه نویسی شده مقایسه و آزمایش می شوند. نتایج تجربی نشان می دهد که این ترکیب کارآمد است و در واقع دقت تشخیص را بهبود می بخشد.

روش پیشنهادی

شکل ۱ چارچوب کلی روش پیشنهادی نشان داده شده است که در آن، مراحل مختلف از پیش‌پردازش تا تشخیص ناهنجاری با ترکیب تطبیقی الگوریتم‌های SVM، SGD-SVM و Isolation Forest، LOF به‌صورت تصویری نمایش داده شده است. به عنوان هدف اصلی طراحی این چارچوب می توان گفت، ترکیب چند الگوریتم بدون نظارت (مثل LOF، Isolation Forest، One-Class SVM، و SVM) صورت تطبیقی، یعنی وزن هر الگوریتم بر اساس عملکردش روی داده‌ها تنظیم شود. یعنی به‌جای استفاده مستقل از الگوریتم‌هایی مثل LOF یا iForest، یک چارچوب تطبیقی طراحی می کنیم که بسته به نوع داده، وزن هر الگوریتم تنظیم شود. این کار باعث افزایش دقت و کاهش نرخ خطا شده و نوآوری محسوب می‌شود چون کمتر کسی از ensemble در یادگیری بدون نظارت استفاده کرده است



شکل ۱- چارچوب کلی روش پیشنهادی

در ادامه هر یک از مراحل موجود در روش پیشنهادی را توضیح می دهیم.

- **مرحله جمع‌آوری داده‌ها**

جمع‌آوری داده‌ها یکی از مراحل بنیادین در هر پژوهش علمی محسوب می‌شود، زیرا کیفیت و ساختار داده‌ها تأثیر مستقیمی بر اعتبار نتایج نهایی دارد. در این پژوهش، داده‌ها از پایگاه داده ۱۹۹۹KDD Cup استخراج شده‌اند که یکی از معتبرترین مجموعه داده‌ها در حوزه تشخیص نفوذ و ناهنجاری در شبکه‌های کامپیوتری است. این پایگاه داده شامل میلیون‌ها رکورد تراکنشی است که هر رکورد نمایانگر یک اتصال شبکه با ویژگی‌هایی مانند مدت زمان، نوع پروتکل، تعداد بایت‌های ارسال‌شده، وضعیت اتصال و غیره می‌باشد. داده‌های KDD به‌صورت خام شامل نمونه‌های عادی و ناهنجار هستند که در قالب چندین کلاس مختلف دسته‌بندی شده‌اند. برای اهداف این پژوهش، تنها از ویژگی‌های عددی استفاده شده و نمونه‌های ناهنجار به‌صورت یک کلاس مجزا در نظر گرفته شده‌اند. همچنین، به‌منظور کاهش حجم داده‌ها و افزایش کارایی مدل، از یک زیرمجموعه متوازن از داده‌ها استفاده شده است که شامل تعداد مساوی از نمونه‌های عادی و ناهنجار می‌باشد.

در مرحله جمع‌آوری، توجه ویژه‌ای به موارد زیر شده است:

- ✓ انتخاب داده‌هایی با تنوع بالا در ویژگی‌ها برای افزایش قابلیت تعمیم مدل.
- ✓ حذف داده‌های تکراری یا ناسازگار که ممکن است باعث بایاس در نتایج شوند.
- ✓ بررسی ساختار داده‌ها از نظر توزیع آماری، همبستگی بین ویژگی‌ها و وجود نقاط پرت.

هدف از این مرحله، فراهم‌سازی بستری مناسب برای اجرای الگوریتم‌های یادگیری ماشین و اطمینان از کیفیت داده‌های ورودی است.

- **پیش‌پردازش داده‌ها**

پیش‌پردازش داده‌ها مرحله‌ای حیاتی در آماده‌سازی داده‌ها برای تحلیل و مدل‌سازی است. داده‌های خام معمولاً شامل نویز، مقادیر گمشده، مقیاس‌های متفاوت و ویژگی‌های نامرتبط هستند که می‌توانند عملکرد مدل را به‌شدت تحت تأثیر قرار دهند. در این پژوهش، مراحل پیش‌پردازش با دقت و بر اساس اصول علمی زیر انجام شده‌اند:

- **پاک‌سازی داده‌ها**

در این مرحله، نمونه‌هایی که دارای مقادیر گمشده، نامعتبر یا خارج از دامنه منطقی بودند، حذف شدند. همچنین، داده‌های تکراری شناسایی و حذف گردیدند تا از بایاس آماری جلوگیری شود.

- **نرمال‌سازی ویژگی‌ها**

با توجه به اینکه الگوریتم‌های یادگیری ماشین نسبت به مقیاس ویژگی‌ها حساس هستند، تمامی ویژگی‌های عددی با استفاده از روش Min-Max Scaling به بازه [۰,۱] نگاشت شدند. این کار باعث می‌شود که ویژگی‌هایی با دامنه بزرگ‌تر تأثیر نامتناسبی بر مدل نداشته باشند (Dongqi, et al. (2021).).

- **رمزگذاری ویژگی‌های غیرعددی**

ویژگی‌هایی مانند نوع پروتکل یا وضعیت اتصال که به‌صورت متنی بودند، با استفاده از روش One-Hot Encoding به فرم عددی تبدیل شدند. این روش باعث می‌شود که مدل بتواند تفاوت بین مقادیر متنی را به‌درستی درک کند بدون اینکه ترتیب خاصی به آن‌ها نسبت دهد.

- **انتخاب ویژگی‌ها**

با استفاده از روش‌های آماری مانند تحلیل همبستگی و آزمون‌های اطلاعات متقابل، ویژگی‌هایی که تأثیر کمی بر تشخیص ناهنجاری داشتند، حذف شدند. این کار باعث کاهش پیچیدگی مدل و افزایش سرعت آموزش شد.

- **تقسیم داده‌ها**

داده‌ها به دو بخش آموزش (۷۰٪) و تست (۳۰٪) تقسیم شدند تا ارزیابی مدل به‌صورت مستقل انجام شود. این تقسیم‌بندی به‌صورت تصادفی و با حفظ نسبت نمونه‌های عادی و ناهنجار انجام شد.

پیش‌پردازش دقیق داده‌ها نه‌تنها باعث افزایش دقت مدل می‌شود، بلکه از بروز خطاهای محاسباتی و نتایج گمراه‌کننده جلوگیری می‌کند.

مرحله اجرای الگوریتم‌های پایه

در این مرحله، الگوریتم‌های پایه‌ای که در چارچوب ترکیبی تطبیقی مورد استفاده قرار می‌گیرند، به‌صورت مستقل و موازی بر روی داده‌های پیش‌پردازش‌شده اجرا می‌شوند. هدف از این مرحله، استخراج نمرات ناهنجاری از دیدگاه‌های مختلف الگوریتمی است تا بتوان در مراحل بعدی آن‌ها را ترکیب و تحلیل کرد. انتخاب الگوریتم‌ها بر اساس معیارهایی مانند تنوع روش‌شناسی، قابلیت تفسیر، سرعت اجرا و عملکرد تجربی در مطالعات پیشین انجام شده است.

این مرحله نقش کلیدی در آماده‌سازی داده‌ها برای تجمیع نهایی دارد و پایه‌ای برای محاسبه وزن‌های تطبیقی فراهم می‌سازد.

مرحله محاسبه وزن‌های تطبیقی

در یک چارچوب ترکیبی، همه الگوریتم‌ها به‌صورت برابر عمل نمی‌کنند. برخی الگوریتم‌ها در تشخیص ناهنجاری‌ها عملکرد بهتری دارند و برخی دیگر ممکن است در شرایط خاص دچار خطا شوند. از این‌رو، استفاده از وزن‌های تطبیقی برای هر الگوریتم ضروری است تا تأثیر آن‌ها در خروجی نهایی متناسب با کیفیت عملکردشان تنظیم شود.

- **معیارهای تعیین وزن**

در این پژوهش، وزن هر الگوریتم بر اساس سه معیار اصلی تعیین شده است:

- ✓ دقت نسبی: اگر داده‌ها دارای برچسب باشند، می‌توان عملکرد هر الگوریتم را با معیارهایی مانند Precision، Recall و F1-Score ارزیابی کرد. الگوریتم‌هایی با دقت بالاتر وزن بیشتری دریافت می‌کنند((Taejin Lee. (2020) .
- ✓ همبستگی خروجی‌ها: با محاسبه ضریب همبستگی پیرسون بین نمرات ناهنجاری الگوریتم‌ها، می‌توان میزان توافق آن‌ها را سنجید. الگوریتم‌هایی که خروجی آن‌ها با سایر الگوریتم‌ها هم‌راستا باشد، قابل اعتمادتر تلقی می‌شوند.
- ✓ پایداری نمرات: با اعمال تغییرات جزئی در داده‌ها (مثلاً حذف تصادفی ۵٪ نمونه‌ها)، نمرات ناهنجاری مجدداً محاسبه می‌شوند. الگوریتم‌هایی که نمراتشان پایدار باقی بماند، وزن بیشتری دریافت می‌کنند.

- **نرمال‌سازی وزن‌ها**

پس از محاسبه وزن‌ها، آن‌ها به‌گونه‌ای نرمال‌سازی می‌شوند که مجموع وزن‌ها برابر با ۱ باشد. این کار باعث می‌شود که در مرحله تجمیع نهایی، سهم هر الگوریتم به‌صورت نسبی و قابل مقایسه لحاظ شود. استفاده از وزن‌های تطبیقی باعث می‌شود که چارچوب ترکیبی نه‌تنها از تنوع الگوریتم‌ها بهره‌مند شود، بلکه به‌صورت هوشمندانه از الگوریتم‌های قوی‌تر بیشتر استفاده کند و تأثیر الگوریتم‌های ضعیف‌تر را کاهش دهد.

تجمیع نهایی نمرات ناهنجاری

پس از استانداردسازی نمرات و تعیین وزن‌های تطبیقی، مرحله نهایی شامل ترکیب نمرات ناهنجاری از الگوریتم‌های مختلف برای هر نمونه است. هدف از این مرحله، تولید یک نمره ترکیبی است که نمایانگر ارزیابی کلی سیستم از میزان غیرعادی بودن هر نمونه باشد. این روش باعث می‌شود که نمره نهایی هر نمونه تحت تأثیر الگوریتم‌هایی قرار گیرد که عملکرد بهتری داشته‌اند.

- تحلیل نمرات ترکیبی

پس از محاسبه نمرات نهایی، توزیع آن‌ها مورد بررسی قرار گرفت. نمودارهای چگالی، هیستوگرام و جعبه‌ای برای تحلیل پراکندگی نمرات استفاده شدند. همچنین، نقاط پرت در نمرات ترکیبی شناسایی شدند تا در مرحله بعدی به‌عنوان ناهنجار علامت‌گذاری شوند.
تجمیع نمرات با استفاده از وزن‌های تطبیقی دارای مزایای زیر است (Ankit, et al. (2021),):

- ✓ افزایش دقت تشخیص ناهنجاری‌ها
- ✓ کاهش تأثیر الگوریتم‌های ضعیف‌تر
- ✓ انعطاف‌پذیری در مواجهه با داده‌های متنوع
- ✓ قابلیت تفسیر بهتر خروجی نهایی

این مرحله نقش کلیدی در موفقیت چارچوب ترکیبی دارد و پایه‌ای برای تصمیم‌گیری نهایی در مورد ناهنجاری‌ها فراهم می‌سازد.

تشخیص ناهنجاری

پس از محاسبه نمرات نهایی برای هر نمونه، مرحله تشخیص ناهنجاری آغاز می‌شود. در این مرحله، هدف آن است که با استفاده از

نمرات ترکیبی حاصل از الگوریتم‌های مختلف، نمونه‌هایی که رفتار غیرعادی دارند شناسایی شوند. این فرآیند شامل تعیین آستانه تصمیم‌گیری، تحلیل آماری نمرات و ارزیابی عملکرد مدل است.

- **تعیین آستانه (Threshold)**

برای تشخیص ناهنجاری، لازم است یک آستانه عددی تعیین شود که نمرات بالاتر از آن به‌عنوان ناهنجار تلقی شوند. در این پژوهش، از روش‌های آماری برای تعیین آستانه استفاده شده است. یکی از روش‌های رایج، استفاده از صدک ۹۵ یا ۹۹ نمرات نهایی است. به‌عبارت دیگر، ۵٪ یا ۱٪ از نمونه‌هایی که بالاترین نمرات را دارند، به‌عنوان ناهنجار شناسایی می‌شوند.
همچنین، در صورت وجود داده‌های برچسب‌خورده، می‌توان از نمودار ROC و محاسبه نقطه تعادل بین نرخ تشخیص صحیح و نرخ خطای مثبت کاذب برای تعیین آستانه بهینه استفاده کرد (Kumari, S. e. (2024)) .

- **تحلیل بصری ناهنجاری‌ها**

برای درک بهتر نحوه تفکیک نمونه‌های عادی و ناهنجار، از روش‌های مجازی‌سازی مانند PCA و t-SNE استفاده شده است. این روش‌ها داده‌ها را به فضای دوبعدی نگاشت می‌کنند تا بتوان توزیع نمونه‌ها را مشاهده کرد. در نمودارهای حاصل، نمونه‌های ناهنجار معمولاً در حاشیه یا خارج از خوشه‌های اصلی قرار دارند.

ارزیابی عملکرد مدل

در صورت دسترسی به برچسب‌های واقعی داده‌ها، عملکرد مدل با استفاده از معیارهای زیر ارزیابی شده است (Wang, (2019)) :

Accuracy: نسبت نمونه‌های ناهنجار درست شناسایی‌شده به کل نمونه‌های شناسایی‌شده به‌عنوان ناهنجار. (فرمول ۱)

فرمول ۱

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: صحت برای نقاط پرت، نسبت مشاهدات مثبت پیش‌بینی‌شده درست (مثبت واقعی) به کل مثبت‌های پیش‌بینی‌شده (مجموع مثبت‌های درست و مثبت کاذب) است (فرمول ۲).

فرمول ۲

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall: نسبت نمونه‌های ناهنجار درست شناسایی‌شده به کل نمونه‌های واقعی ناهنجار.(فرمول ۳)

فرمول ۳

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-Score میانگین هماهنگ Precision و Recall. (فرمول ۴)

فرمول ۴

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

نتایج حاصل از ارزیابی نشان می‌دهد که چارچوب ترکیبی تطبیقی عملکرد بهتری نسبت به الگوریتم‌های منفرد دارد و توانایی بالایی در شناسایی ناهنجاری‌ها از دیدگاه‌های مختلف دارد.

در فرمولهای بالا داریم:

TP (مثبت واقعی): تعداد نقاط پرت که به درستی به عنوان نقاط پرت شناسایی شده اند.

TN (منفی های واقعی): تعداد مشاهدات عادی که به درستی به عنوان عادی شناسایی شده اند.

FP (مثبت کاذب): تعداد مشاهدات عادی که به اشتباه به عنوان نقاط پرت شناسایی شده اند.

FN (منفی کاذب): تعداد نقاط پرت که به اشتباه به عنوان مشاهدات عادی شناسایی شده اند.

تجزیه و تحلیل مقایسه ای، شامل تجزیه و تحلیل مقایسه ای مدل ها، بحث در مورد نقاط قوت و ضعف آنها در تشخیص ناهنجاری ها است. این تحلیل توسط نمایش‌های بصری و معیارهای آماری پشتیبانی می‌شود که مناسب بودن هر الگوریتم برای انواع مختلف ناهنجاری‌های داده را برجسته می‌کند.

پیاده سازی و ارزیابی کارایی روش پیشنهادی

مراحل طی شده برای پیاده سازی و ارزیابی کارایی روش پیشنهادی عبارتند از:

- جمع آوری داده ها

داده های مورد نیاز برای پیاده سازی روش پیشنهادی از پایگاه داده ۹۹KDD جمع آوری خواهد شد. پایگاه داده ۹۹KDD عمدتاً برای تشخیص نفوذ در شبکه طراحی شده است و شامل داده‌های مربوط به حملات سایبری و فعالیت‌های مشکوک در شبکه‌های کامپیوتری است. این مجموعه داده برای ارزیابی سیستم‌های تشخیص نفوذ (IDS) استفاده می‌شود و شامل اطلاعاتی درباره انواع حملات و ارتباطات شبکه‌ای است. مجموعه داده‌هایی مانند Bank Transaction Dataset for Fraud Detection در Kaggle وجود دارند که شامل اطلاعات تراکنش‌های بانکی برای تشخیص تقلب هستند. لذا از مجموع این مجموعه داده ها برای این پژوهش استفاده خواهیم کرد.. مجموعه داده Bank Transaction Dataset for Fraud Detection در Kaggle شامل اطلاعات تراکنش‌های مالی برای تحلیل و تشخیص تقلب است. این مجموعه داده شامل ویژگی‌هایی مانند مقدار تراکنش، شناسه مشتری، نوع تراکنش، زمان انجام تراکنش، و وضعیت تقلبی بودن یا نبودن است. در Kaggle مجموعه داده‌های مختلفی برای تراکنش‌های بانکی موجود است. یکی از این مجموعه داده‌ها شامل ۱۰۰۰ تراکنش بانکی است که شامل اطلاعاتی مانند شناسه تراکنش، مبلغ، نوع تراکنش (انتقال، برداشت، واریز)، زمان انجام تراکنش، وضعیت تراکنش (موفق یا ناموفق)، پرچم تقلب، موقعیت جغرافیایی، دستگاه مورد استفاده، تأخیر شبکه و پهنای باند می‌شود. جدول ۱ ویژگیهای موجود در مجموعه داده های جمع آوری شده را نشان می دهد.

جدول ۱-ویژگیهای موجود در مجموعه داده های جمع آوری شده[۱۶]

ویژگی	توضیحات
Transaction Amount	مقدار پولی که در جریان تراکنش، توسط مشتری جابجا شده است.
Transaction Date	تاریخ انجام تراکنش

TransactionType	نوع تراکنش مانند خرید آنلاین، برداشت از طریق ATM، یا انتقال از طریق موبایل بانک یا اینترنت بانک و غیره.
Card holder Name	نام دارنده کارت بانکی (مشتری بانک)
Card number	شماره کارتی که با استفاده از آن مشتری اقدام به انجام تراکنش می کند.
Card Expire Date	آخرین مدت زمانی که شخص می تواند از کارت اعتباری استفاده کند.
Merchant Name	نام تاجر یا شرکت تجاری که معامله در آن انجام شده است.
Merchant Location	موقعیت مکانی (آدرس، شهر، کشور) فروشنده
IP Address	آدرس IP مرتبط با تراکنش
Device Information	اطلاعات مربوط به دستگاه مورد استفاده برای تراکنش، مانند نوع دستگاه، سیستم عامل و مرورگر
Geolocation	موقعیت جغرافیایی تراکنش بر اساس مختصات GPS یا سایر داده‌های مکانی
Transaction Status	وضعیت تراکنش، مانند تایید شده، رد شده یا در انتظار
Fraud Flag	یک پرچم دودویی که نشان می‌دهد آیا تراکنش به عنوان یک تراکنش بالقوه جعلی علامت‌گذاری شده است یا خیر؟
Reason Code	کدی که دلیل علامت‌گذاری تراکنش به عنوان تراکنش بالقوه جعلی را نشان می‌دهد.
User Profile	اطلاعات مربوط به حساب کاربری، سابقه و رفتار کاربر، مانند سن حساب، تعداد تراکنش‌ها و الگوهای خرج کردن
Risk Score	امتیاز عددی که سطح ریسک مرتبط با تراکنش را نشان می‌دهد.
Authentication Method	روشی که برای تأیید تراکنش استفاده می‌شود، مانند پین، رمز عبور یا تأیید بیومتریک
Response Time	زمان صرف شده برای پاسخ به تراکنش، شامل فرآیندهای مجوز و تأیید
Linked Accounts	اطلاعات مربوط به سایر حساب‌های مرتبط با حساب کاربر، مانند حساب‌های مشترک یا کاربران مجاز
Previous Fraud History	هرگونه نمونه قبلی کلاهبرداری یا فعالیت مشکوک مرتبط با کاربر یا کارت مورد استفاده در تراکنش

برای پیاده سازی روش پیشنهادی، از کل مجموعه داده های جمع آوری شده، با استفاده از روش نمونه برداری ساده تصادفی، ۷۰ درصد از داده ها را به عنوان مجموعه داده های تمرین و ۳۰ درصد را به عنوان مجموعه داده های تست انتخاب می کنیم.

- پردازش داده ها

برای پردازش داده‌های تراکنش‌های بانکی در پایتون، مراحل زیر را طی می کنیم:

- ✓ بارگذاری داده‌ها
- ✓ بررسی و پاک‌سازی داده‌ها
- ✓ نرمال‌سازی و مقیاس‌بندی داده‌ها

- اجرای الگوریتم های پیشنهادی بر روی مجموعه داده ها

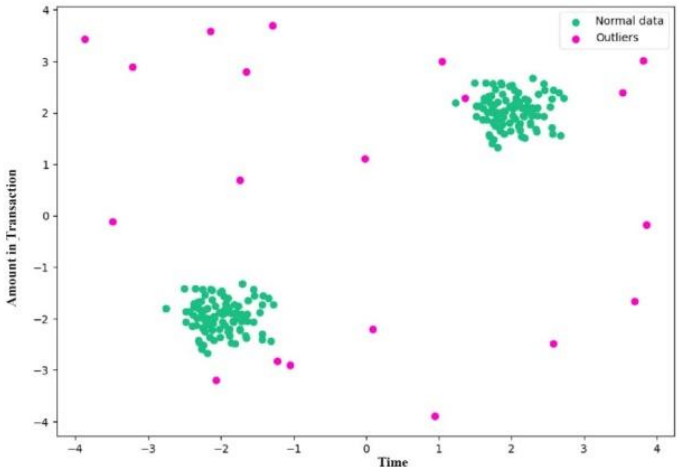
تشخیص موارد خارج از قاعده در داده‌ها (داده های پرت یا ناهنجاری ها) یکی از جنبه‌های مهم تحلیل داده است و روش‌های مختلفی برای این کار وجود دارد. برخی از رایج‌ترین روش‌ها عبارتند از (Obeng, S. e. (2024).):

- ✓ روش‌های آماری
- ✓ روش‌های مبتنی بر یادگیری ماشین

در این تحقیق از روشهای یادگیری نظارت نشده (شامل ماشین بردار ماشین یک کلاسه، SVM یک کلاسی با نزول گرادیان تصادفی (SGD)، جنگل جداسازی (iForest) وعامل دورافتاده محلی (LOF)) استفاده می کنیم.

شکل ۲، یک نمودار پراکندگی داده‌های مورد استفاده در این پژوهش را نشان می‌دهد که دارای دو خوشه اصلی از داده‌های نرمال (نقاط سبز) و چندین نقاط پرت (صورتی) است. محورهای نمودار با عنوان میزان پول جابجا شده در تراکنشها (به عنوان ویژگی شماره ۱) و زمان انجام تراکنشها (به عنوان ویژگی شماره ۲) برچسب‌گذاری شده‌اند و مقادیر آن‌ها بین -۴ تا ۴ متغیر است. علت اینکه مقادیر ویژگیها بین ۴ تا -۴ تنظیم شده است، این است که استانداردسازی داده ها انجام شده و این بازه ها در نتیجه اعمال

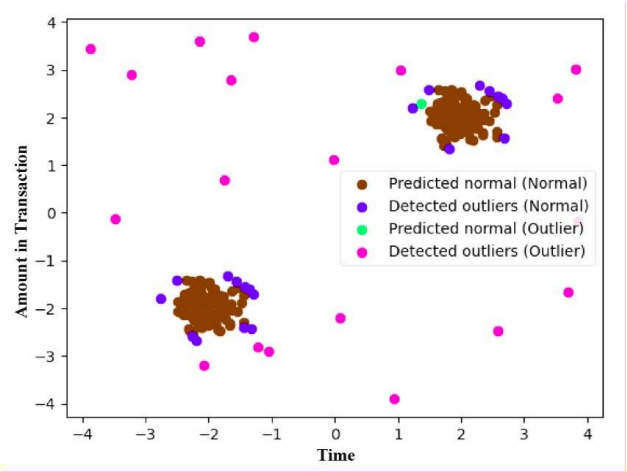
تکنیک Z-Score Normalization انتخاب شده است که باعث می شود تا مقادیر داده‌ها را در یک محدوده استاندارد قرار می‌دهد تا مدل‌های آماری بهتر عمل کنند. در نتیجه تمام داده هایی که در این بازه [-۴, ۴] قرار می گیرند به حالت داده های نرمال در نظر گرفته شده و داده های خارج از این محدوده به‌عنوان نقاط پرت شناخته می‌شوند. مقیاس -۴ تا ۴ در این تحقیق، انتخاب شده است تا داده‌های نرمال و نقاط پرت به وضوح قابل مشاهده باشند و تحلیل آن‌ها راحت‌تر باشد. از طرفی، استفاده از بازه‌ی محدود باعث می‌شود که محاسبات کوواریانس و شناسایی نقاط پرت مؤثرتر انجام شود. لازم به ذکر است که، این بازه‌ی عددی بسته به روش تولید داده‌ها و هدف تحقیق ممکن است متفاوت باشد. شکل ۲، تصویر یک نمودار پراکندگی را نشان می‌دهد که داده‌ها را به دو گروه "داده‌های عادی" و "موارد خارج از قاعده" دسته‌بندی می‌کند. محور افقی برچسب "زمان" دارد، و محور عمودی نشان‌دهنده "مقدار تراکنش" است. این نمودار مشخص می‌کند که نقاط سبز نمایانگر "داده‌های عادی" هستند، و نقاط صورتی نشان‌دهنده "موارد خارج از قاعده" می‌باشند. نقاط سبز به‌طور عمده در دو گروه مجزا جمع شده‌اند: یکی در اطراف (۱,۲) و دیگری در اطراف (-۲,-۲). در مقابل، نقاط صورتی پراکنده شده‌اند که نشان می‌دهد این موارد از الگوی معمولی خارج هستند. این نوع تصویر برای تحلیل تراکنش‌ها و شناسایی موارد غیرعادی بسیار مفید است.



شکل ۲-نمودار پراکندگی داده ها بر اساس دو ویژگی تعیین شده

- **پیاده‌سازی الگوریتم‌ها جنگل جداسازی iForest**

شکل ۳ نتایج تشخیص ناهنجاری با استفاده از iForest را روی مجموعه داده‌های جمع آوری شده در این تحقیق نشان می‌دهد. این روش به جای ایجاد مرز برای جدا کردن ناهنجاری‌ها از مشاهدات عادی، آنها را جدا می‌کند.



شکل ۳- نتایج تشخیص ناهنجاری با اعمال الگوریتم iForest

در شکل ۳، نمودار دو گروه اصلی نقاط قهوه‌ای را نشان می‌دهد که حاکی از تراکنش‌های عادی است. در میان این نقاط، تعدادی نقاط بنفش دیده می‌شود که نشان‌دهنده‌ی نابهنجاری‌های شناسایی‌شده در میان تراکنش‌های عادی هستند. همچنین، نقاط صورتی پراکنده در نمودار نشان‌دهنده‌ی تراکنش‌هایی هستند که به‌عنوان نابهنجاری تشخیص داده شده‌اند. این نمودار به ارزیابی عملکرد الگوریتم تشخیص نابهنجاری کمک می‌کند. در زمینه‌هایی مانند شناسایی تراکنش‌های مشکوک یا تقلبی در سیستم‌های مالی بسیار مفید است.

- پیاده‌سازی ماشین بردار پشتیبان یک کلاسه (One-Class SVM)

نتایج این پیاده سازی نشان می دهد که، SVM تک‌کلاسه با موفقیت اکثر نقاط داده نرمال را به درستی طبقه‌بندی کرده است و این نتیجه نشان دهنده توانایی مدل در تمایز بین داده‌های نرمال و ناهنجاری‌ها در یک مجموعه داده است و آن را به ابزاری مفید برای وظایف تشخیص ناهنجاری تبدیل می‌کند.

- پیاده سازی عامل دورافتاده محلی (LOF)

با توجه به نتایج بدست آمده از این پیاده سازی، روش LOF، انحراف چگالی محلی یک نقطه داده معین را نسبت به همسایگانش ارزیابی می‌کند و هدف آن شناسایی مناطقی با چگالی مشابه و برجسته کردن نقاطی است که به عنوان ناهنجاری برجسته می‌شوند. روش LOF با تمرکز بر همسایگی محلی هر نقطه داده، به طور مؤثر داده‌های پرت را شناسایی می‌کند. این تکنیک به ویژه در مجموعه داده‌هایی که چگالی اطراف مشاهدات عادی و ناهنجاری‌ها به طور قابل توجهی متفاوت است، مفید است. همانطور که نشان داده شده است، مدل LOF با موفقیت داده‌های پرت مصنوعی حداقلی را به عنوان داده‌های پرت علامت‌گذاری کرده است .(Cholevas, C. e. (2024).)

- **پیاده سازی مرحله استانداردسازی نمرات ناهنجاری**

پیاده‌سازی مرحله استانداردسازی نمرات ناهنجاری را با استفاده از زبان پایتون و کتابخانه‌های رایج مانند NumPy و Pandas انجام می دهیم. این مرحله معمولاً پس از اجرای الگوریتم‌های تشخیص ناهنجاری انجام می‌شود تا نمرات خروجی آن‌ها قابل مقایسه و ترکیب باشند.

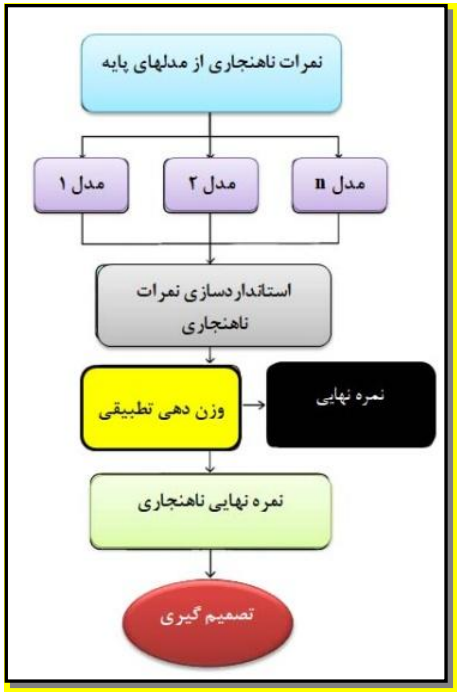
- **پیاده سازی مرحله محاسبه وزن های تطبیقی**

هدف از این مرحله، تخصیص وزن به هر الگوریتم بر اساس میزان توافق آن با سایر الگوریتم‌ها یا بر اساس توانایی‌اش در شناسایی ناهنجاری‌ها است. روش پیشنهادی برای این منظور، روش توافق مبتنی بر همبستگی است. در این روش، وزن هر الگوریتم بر اساس میانگین همبستگی آن با سایر الگوریتم‌ها محاسبه می‌شود. الگوریتم‌هایی که خروجی‌شان با سایر الگوریتم‌ها هم‌راستاتر باشد، وزن بیشتری می‌گیرند. این وزن‌ها نشان می‌دهند که هر الگوریتم چه سهمی در نمره نهایی دارد. اگر الگوریتمی خروجی متفاوت‌تری داشته باشد، وزنش کمتر خواهد بود.

- پیاده سازی مرحله تجمیع نمرات ناهنجاری

نمودار تصویری فرآیند ترکیب نمرات ناهنجاری در مدل ترکیبی پیشنهادی در شکل ۴ مراحل زیر را به‌صورت گرافیکی نمایش می‌دهد.

- ✓ دریافت نمرات ناهنجاری از چند مدل پایه (مثل SVM، Isolation Forest، LOF)
- ✓ استانداردسازی نمرات برای هم‌مقیاس‌سازی
- ✓ اعمال وزن‌های تطبیقی بر اساس عملکرد هر مدل
- ✓ محاسبه نمره نهایی با فرمول میانگین وزنی
- ✓ تصمیم‌گیری نهایی بر اساس نمره ترکیبی



شکل ۴-فرایند ترکیب نمرات ناهنجاری

- پیاده سازی مرحله تشخیص ناهنجاری: تشخیص ناهنجاری، یعنی تعیین اینکه کدام نمونه‌ها واقعاً ناهنجار هستند. این مرحله شامل تعیین آستانه (Threshold) و برچسب‌گذاری (Labeling) است. یکی از روش‌های رایج برای تعیین آستانه، استفاده از درصدی از داده‌هاست.

- **ارزیابی کارایی روش پیشنهادی**

جدول ۲ نتایج حاصل از ارزیابی کارایی روش پیشنهادی را نشان می دهد.

جدول ۲-نتایج ارزیابی کارایی روش پیشنهادی

مدل	دقت	صحت	فراخوانی	F۱ Score
SVM تک کلاسه	٪۹۰.۹	٪۵۰.۱	٪۹۹.۹	٪۶۶.۷
SVM تک کلاسه با SGD	٪۹۱.۴	٪۹۹.۹	٪۵۰.۱	٪۹۵.۲
جنگل جداسازی	٪۹۰.۵	٪۴۸.۷	٪۹۵.۱	٪۶۴.۴
LOF	٪۸۲.۷	٪۹۰.۱	٪۹۹.۱	٪۹۵.۲
مدل ترکیبی	٪۹۹.۶	٪۹۹.۹	٪۹۹.۹	٪۹۸.۹

با توجه به جدول ۲ می توان نتیجه گیری کرد که، مدل‌های SVM تک‌کلاسه، دقت و صحت یکسانی را نشان می‌دهند و هر دو با موفقیت تمام داده‌های پرت را به خاطر می‌آورند که منجر به امتیاز F1 تقریباً ٪۶۶.۷ می‌شود. نکته قابل توجه این است که SVM تک‌کلاسه با SGD به بالاترین دقت با ٪۹۱.۴ دست می‌یابد که نشان می‌دهد مشاهدات عادی را به اشتباه به عنوان داده‌های پرت برچسب‌گذاری نمی‌کند. با این حال، کمترین میزان فراخوانی ٪۵۰.۱ را دارد که نشان می‌دهد در شناسایی اکثر داده‌های پرت واقعی شکست می‌خورد، که این امر در امتیاز F1 پایین 95.2٪ آن نیز منعکس شده است. مدل جنگل جداسازی، یک پروفایل متعادل با دقت مناسب (٪۹۰.۵) و دومین میزان فراخوانی بالا (٪۹۵.۱) را نشان می‌دهد که منجر به امتیازF1 برابر با ٪۶۴.۴ می‌شود. مدل «فاکتور داده‌های پرت محلی» در این مقایسه با کمترین دقت (٪۸۲.۷) و امتیازF1 برابر با ٪۹۵.۲ (در کنار یک «بازخوانی» ماقبل آخر نسبت به «ماشین بردار پشتیبان تک‌کلاسه» باSGD ، با مشکل مواجه است.

ماشین بردار پشتیبان تک‌کلاسه و «کوواریانس قوی» عملکرد متعادلی با امتیازهای برابر در بازخوانی و F۱(به ترتیب ٪۹۹.۱ و ٪۶۶.۶۷) نشان می‌دهند که نشان‌دهنده عملکرد قوی در تشخیص داده‌های پرت واقعی است. با این حال، دقت آنها متوسط است، که نشان می‌دهد برخی از نقاط عادی به اشتباه به عنوان ناهنجاری برچسب‌گذاری شده‌اند. «ماشین بردار پشتیبان تک‌کلاسه» با SGD دقت بالایی (٪۱۰۰.۰۰) اما امتیاز بازخوانی و F1 بسیار پایینی (به ترتیب ٪۵۰ و ٪۹۵.۲) نشان می‌دهد، که رویکرد محافظه‌کارانه آن را در برچسب‌گذاری داده‌های پرت برجسته می‌کند. این مدل بسیاری از داده‌های پرت واقعی را از دست می‌دهد، اما وقتی یک نقطه را به عنوان داده پرت برچسب‌گذاری می‌کند، بسیار دقیق است. جنگل جداسازی تعادل خوبی بین بازخوانی و دقت ایجاد می‌کند و به امتیاز F1 برابر با ٪۶۴.۴ دست می‌یابد. این نتایج نشان دهنده توانایی قوی در شناسایی نقاط پرت ضمن حفظ نرخ معقولی از تشخیص‌های کاذب است. عامل پرت محلی کمترین عملکرد را در بین تمام معیارها نشان می‌دهد که نشان‌دهنده دشواری در تشخیص دقیق بین نقاط نرمال و نقاط پرت در این مجموعه داده است. همانطور که مشاهده می شود در روش ترکیبی همه معیارهای عملکردی دارای درصد بالاتری نسبت به تک تک مدلهای استفاده شده به تنهایی می باشد. جدول 3 تجزیه و تحلیل مقایسه‌ای از کارایی محاسباتی الگوریتمهای تشخیص ناهنجاری بدون نظارت استفاده شده در این تحقیق را نشان می‌دهد.

جدول 3- تجزیه و تحلیل مقایسه‌ای از کارایی محاسباتی الگوریتمهای تشخیص ناهنجاری بدون نظارت استفاده شده در این پژوهش

مدل	زمان برازش مدل	زمان پیش بینی مدل	زمان رسم نتایج
SVM تک کلاسه	۰,۰۰۲	۰,۰۰۰۱	۱,۱۲۵
SVM تک کلاسه با SGD	۰,۰۰۲	۰,۰۰۰۱	۰,۴۵۶
جنگل جداسازی	۰,۲۰۴	۰,۰۰۴۰	۰,۴۰۵
LOF	۰,۰۰۸	۰,۳۰۱۲	۰,۰۵۰
روش ترکیبی	۰,۰۰۴	۰,۰۰۰۰۸	۰,۰۱

با توجه به جدول ۳ می توان استدلال کرد که، SVM تک کلاسه و SVM تک کلاسه با SGD سریع‌ترین زمان برازش مدل و پیش‌بینی داده‌های پرت را نشان می‌دهند، هر دو تنها ۰.۰۰۲ ثانیه برای برازش مدل و ۰.۰۰۰۱ ثانیه برای پیش‌بینی داده‌های پرت زمان می‌برند که نشان‌دهنده کارایی بالا در محیط‌های محاسباتی ساده‌تر است. نکته قابل توجه این است که هر دو الگوریتم زمان‌های بسیار متفاوتی برای ترسیم نتایج دارند، به طوری که SVM تک کلاسه در مقایسه با همتای SGD خود (۰.۴۵۶ ثانیه) به طور قابل توجهی طولانی‌تر (۱.۱۲۵ ثانیه) طول می‌کشد. جنگل جداسازی، در حالی که در برازش مدل (۰.۲۰۴ ثانیه) و پیش‌بینی داده‌های پرت (۰.۰۰۴۰ ثانیه) کندتر است، به زمان کمتری برای پیاده سازی نتایج (۰.۴۰۵ ثانیه) نیز نیاز دارد. ضریب داده‌های پرت محلی، زمان برازش مدل متوسطی (۰.۰۰۸ ثانیه) را نشان می‌دهد، اما زمان پیش‌بینی داده‌های پرت نسبتاً بالایی (۰.۳۰۱۲ ثانیه) دارد که نشان‌دهنده‌ی یک تعامل احتمالی بین سرعت برازش و پیچیدگی پیش‌بینی است. کوواریانس قوی، اگرچه کندترین زمان برازش مدل (۰.۰۶۰ ثانیه) را دارد، اما پیش‌بینی داده‌های پرت کارآمد (۰.۰۰۰۹ ثانیه) و سریع‌ترین ترسیم نتایج (۰.۰۱۴ ثانیه) را نشان می‌دهد. این تحلیل کارایی محاسباتی، عملکرد متفاوت تعامل بین مدلهای مختلف تشخیص ناهنجاری را از نظر هزینه محاسباتی و کارایی برجسته می‌کند و بینش‌های ارزشمندی را برای انتخاب مدل‌های مناسب بر اساس الزامات کاربردی خاص ارائه می‌دهد. با توجه به جدول 5، روش ترکیبی از نظر زمان برازش مدل نسبت به مدلهای SVM تک کلاسه و مدلهای SVM تک کلاسه با SGD بیشتر است ولی از نظر پیش بینی مدل و زمان رسم نتایج در مقایسه با تک تک مدلها، بهبود یافته است و مدت زمان لازم برای پیش بینی مدل و رسم نتایج، در روش ترکیبی کاهش یافته است.

کارایی روش های مورد استفاده در این تحقیق را با روشهای پیشنهادی توسط محققان دیگر [Kumari, S. e. (2024)., Bajpai, A. D. (2024)., Obeng, S. e. (2024).] از لحاظ معیارهای ارزیابی (شامل دقت، صحت و F1-Score) مورد مقایسه قرار می دهیم. جدول 4 نتایج بدست آمده را نشان می دهد.

جدول ۴-مقایسه کارایی روش پیشنهادی با روشهای دیگر

مدل	دقت	صحت	فراخوانی	F۱ Score
SVM تک کلاسه	٪۹۰,۹	٪۵۰,۱	٪۹۹,۹	٪۶۶,۷
SVM تک کلاسه با SGD	٪۹۱,۴	٪۹۹,۹	٪۵۰,۱	٪۹۵,۲
جنگل جداسازی	٪۹۰,۵	٪۴۸,۷	٪۹۵,۱	٪۶۴,۴
LOF	٪۸۲,۷	٪۹۰,۱	٪۹۹,۱	٪۹۵,۲
مدل ترکیبی	٪۹۹,۶	٪۹۹,۹	٪۹۹,۹	٪۹۸,۹
روش مبتنی بر یادگیری عمیق [۲۳]	٪۸۹,۹	٪۹۱,۷	٪۹۹,۳	٪۹۴,۱
فاکتور پرت محلی (LOF) و رمزگذار خودکار[۲۴]	٪۸۸,۷	٪۸۵,۶	٪۸۴,۲	٪۸۶,۱
روش مبتنی بر فاصله[۲۵]	٪۷۱,۶	٪۶۹,۸	٪۶۸,۱	٪۶۹,۳

لذا با توجه به نتایج مندرج در جدول۴، می توان گفت:

- ✓ روش مبتنی بر یادگیری عمیق ([Obeng, S. e. (2024).]) با تعادل عالی بین دقت و صحت به عنوان مدل خوبی محسوب می شود.
- ✓ مدل‌های متعادل: LOF کوراربائس قوی و SVM تک کلاسه با SGD گزینه‌های مناسبی برای داده‌های با توزیع نرمال

یا پیچیده هستند. در حالیکه روش ترکیبی پیشنهادی بهترین گزینه برای تمام داده ها به خصوص برای داده های با توزیع پیچیده می باشد.

- ✓ مدل‌های ضعیف شامل روش مبتنی بر فاصله ([Kumari, S. e. (2024).]) و جنگل جداسازی به دلیل صحت پایین، برای کاربردهای حساس توصیه نمی‌شوند.

نتیجه و جمع بندی

روش پیشنهادی در این تحقیق، برای تشخیص ناهنجاری در تراکنشهای انجام شده را بر روی انواع داده های تراکنشی (به عنوان مثال در بانکها، شبکه های ارتباطی، شبکه های اجتماعی و غیره) اعمال نموده و اقدام به شناسایی موارد ناهنجاری نمود و بدین وسیله کمک شایسته ای به افزایش امنیت داده ها و جلب رضایت و اعتماد مشتریان نمود. روش پیشنهادی با استفاده از داده های جمع آوری شده از پایگاه داده ۹۹KDD که مربوط به تراکنشهای انجام شده توسط مشتریان یک بانک می باشد، پیاده سازی می شود. هر یک از پنج الگوریتم به طور انحصاری بر روی نقاط داده معمولی آموزش داده می شود. این مرحله بسیار مهم است زیرا به الگوریتم‌ها اجازه می‌دهد تا منطقه فضای ویژگی اشغال شده توسط کلاس اکثریت را بدون تأثیرپذیری از نقاط پرت یاد بگیرند. روش پیشنهادی با استفاده از مدل‌های آموزش‌دیده، ناهنجاری‌ها را با طبقه‌بندی هر نقطه داده در مجموعه داده توسعه‌یافته (از جمله نقاط پرت) به عنوان عادی یا پرت پیش‌بینی می‌کند. این مرحله توانایی هر مدل را برای تعمیم داده‌های آموزشی به داده‌های دیده نشده آزمایش می‌کند. ارزیابی بر اساس دقت، صحت، فراخوانی و امتیاز F1، قابلیت‌های متنوع این الگوریتم‌ها را در شناسایی داده‌های پرت، که هر کدام تحت تأثیر توزیع داده‌های اساسی، پیچیدگی الگوریتم و انتخاب پارامتر هستند، برجسته می‌کند.

مراجع

Bajpai, A. D. (2024). "A novel methodology for anomaly detection in smart home networks via Fractional Stochastic Gradient Descent." ,. Computers and Electrical Engineering 119, 109604.

Bakumenko, A. a. (2022). "Detecting anomalies in financial data using machine learning algorithms." . *Systems 10.5*,, 130.

Ball, R. S. (2023.). A machine learning-based framework for anomaly detection,. *Diss. North-West University (South Africa)*, .

Bolton, R. J. (2021). "Unsupervised profiling methods for fraud detection.",. *Credit scoring and credit control VII*,, 235-255.

Carcillo, F. e. (2021). "Combining unsupervised and supervised learning in credit card fraud detection." ,. Information sciences 557, 317-331.

Cholevas, C. e. (2024). "Anomaly Detection in Blockchain Networks Using Unsupervised Learning: A Survey.",. *Algorithms 17.5* , 201.

Devavarapu, Y. e. (2024). "Credit Card Fraud Detection Using Outlier Analysis and Detection." ,. *4th International Conference on Intelligent Technologies (CONIT). IEEE*, .

Gogoi, P. B. (2020). "Anomaly detection analysis of intrusion data using supervised & unsupervised approach." ,. *J. Convergence Inf. Technol. 5.1*,, 95-110.

Han, Dongqi, et al. (2021). "Deepaid: Interpreting and improving deep learning-based anomaly detection in security applications." *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security.*

Huang, D. e. (2018). "CoDetect: Financial fraud detection with anomaly feature detection." ,. *IEEE Access 6*,, 19161-19174.

Kim, Sujeong, Chanwoong Hwang, and Taejin Lee. (2020), "Anomaly based unknown intrusion detection in endpoint environments." *Electronics 9.6: 1022.*

Kumar, Ankit, et al. (2021), "Distance based pattern driven mining for outlier detection in high dimensional big dataset." *ACM Transactions on Management Information System (TMIS) 13.1 (2021): 1-17.*

Kumari, S. e. (2024). "A Comprehensive Investigation of Anomaly Detection Methods in Deep Learning and Machine Learning: 2019–2023." , IET Information Security 2024.1, 8821891.

Minasyan, Arshak, and Nikita Zhivotovskiy. "Statistically optimal robust mean and covariance estimation for anisotropic Gaussians." arXiv preprint arXiv: 2301.09024 (2023).

Munir, M. e. (2018). "DeepAnT: A deep learning approach for unsupervised anomaly detection in time series.",. *Ieee Access 7*,, 1991-.2005.

Obeng, S. e. (2024). "Utilizing machine learning algorithms to prevent financial fraud and ensure transaction security.", World Journal of Advanced Research and Reviews 23.1,, 1972-1980.

Schlör, D. (2022). Detecting Anomalies in Transaction Data . *(Doctoral dissertation, Universität Würzburg).*

Wang, Hongzhi, Mohamed Jaward Bah, and Mohamed Hammad. "Progress in outlier detection techniques: A survey." IEEE Access 7 (2019): 107964-108000.