DOI: http://dx.doi.org/10.22059/jitm.2016.57818

مدیریت فناوری اطلاعات
Journal of Information Technology Management

دانشکدهٔ مدیریت دانشگاه تهران
دورهٔ ۸، شمارهٔ ۳
پاییز ۱۳۹۵
ص. ۴۹۸- ۴۷۷

کشف تقلب در تراکنش‌های کارت‌های بانکی با استفاده از پردازش موازی ناهنجاری در بزرگ‌داده

محمدرضا تقوا[۱]، طاها منصوری[۲]، کامران فیضی[۳]، بابک اخگر[۴]

چکیده: با رشد روزافزون استفاده از کارت‌های الکترونیکی، به‌خصوص در صنعت بانکی، حجم تراکنش با این کارت‌ها نیز به‌سرعت افزایش پیدا کرده است. به‌علاوه، ذات مالی این کارت‌ها سبب ایجاد مطلوبیت تقلب در این حوزه شده است. تحقیق حاضر با رویکرد پردازش موازی و راه‌حل نگاشت کاهش، از شبکهٔ عصبی مدل کوهونن برای کشف ناهنجاری در تراکنش کارت‌های بانکی استفاده کرده است. برای این منظور، در مرحلهٔ نخست راه‌حلی برای طبقه‌بندی تراکنش‌ها به تقلب‌آمیز و قانونی پیشنهاد شد که نسبت به روش‌های دیگر عملکرد بهتری از خود نشان داد. در مرحلهٔ بعد، روش پیشنهادی به‌دست‌آمده از تبدیل شبکهٔ کوهونن به فرم استفاده‌شدهٔ نگاشت کاهش، توانست قابلیت مناسبی را از نظر زمان اجرا به نمایش بگذارد؛ به‌طوری‌که انتظار می‌رود در تراکنش‌هایی با مفروضات بزرگ‌داده به‌خوبی پیاده‌سازی شود.

واژه‌های کلیدی: بزرگ‌داده، کارت‌های بانکی، کشف تقلب، مدل شبکهٔ عصبی کوهونن.

۱. دانشیار گروه مدیریت صنعتی، دانشکدهٔ مدیریت و حسابداری دانشگاه علامه طباطبایی، تهران، ایران
۲. دانشجوی دکتری مدیریت فناوری اطلاعات، دانشکدهٔ مدیریت و حسابداری دانشگاه علامه طباطبایی، تهران، ایران
۳. استاد گروه مدیریت صنعتی دانشکدهٔ مدیریت و حسابداری دانشگاه علامه طباطبایی، تهران، ایران
۴. استاد گروه انفورماتیک، دانشگاه شفیلد هلم، شفیلد، انگلستان

تاریخ دریافت مقاله: ۱۳۹۵/۰۳/۰۹
تاریخ پذیرش نهایی مقاله: ۱۳۹۵/۰۶/۰۷
نویسندهٔ مسئول مقاله: محمدرضا تقوا
E-mail: Taghva@gmail.com

مقدمه

رشد سریع و پیشرفت در حوزهٔ تجارت الکترونیک، کاربرد کارت‌های بانکی را به‌عنوان ابزاری کارا برای انجام تعاملات الکترونیک افزایش داده است. در کنار همین رشد، تراکنش‌های تقلب‌آمیز نیز به‌صورت روزافزونی در حال افزایش است (میشرا، پاندا، و میشرا، ۲۰۱۳). تقلب کارت بانکی موضوعی حیاتی است که هزینه‌های شایان توجهی برای بانک‌ها و مؤسسه‌های صادرکنندهٔ کارت به‌دنبال دارد. سیستم‌های کشف تقلب با قابلیت تمییزدادن داده‌های تقلب‌آمیز از داده‌های قانونی و نمایان‌کردن رفتارهای متقلبانه همراه با ایجاد قابلیت توسعهٔ راهبردهای مناسب، در کاهش تأثیر آن برای تصمیم‌گیرندگان و کسب‌وکارها بسیار حیاتی هستند (گای، هو، ونگ، چن و سان، ۲۰۱۱).

از آنجا که به‌راحتی نمی‌توان درخصوص انگیزهٔ قانونی یا غیرقانونی‌بودن تراکنش‌ها حکم صادر کرد، بهترین و کم‌هزینه‌ترین ایده، رهگیری تقلب با استفاده از روش‌های ریاضی از میان داده‌های موجود است (فوآ، لی، اسمیت و گایلر، ۲۰۰۵). بنابراین، باید به‌کمک الگوریتم‌های داده‌کاوی بومی‌سازی‌شده، پایگاه‌های داده‌ای بزرگ را تحلیل کرد. با توجه به اهمیت تحلیل بزرگ‌داده برای صنایع مختلف و رشد و توسعهٔ ابزارهای تجارت الکترونیک، به‌خصوص در حوزهٔ کارت‌های بانکی، در ایران نیز حجم انبوهی از داده‌ها جمع‌آوری شده است و استخراج ارزش نهفته در این داده‌ها مانند سایر نقاط دنیا بسیار اهمیت دارد. از سوی دیگر، همان‌طور که پیش‌تر اشاره شد، مسئلهٔ تقلب در این تراکنش‌ها بسیار جدی است؛ بنابراین هدف تحقیق حاضر یافتن مدل مناسب تحلیل داده‌های بزرگ از تراکنش کارت‌های بانکی، برای دستیابی به الگوهایی است که بتوان از طریق آنها تراکنش ورودی را به تراکنش قانونی یا تقلب‌آمیز طبقه‌بندی کرد.

بیان مسئله

در کشورهای دیگر، تحقیقات گسترده‌ای در خصوص استفاده از کارت‌های اعتباری صورت پذیرفته است، اما در ایران استفادهٔ اعتباری از این کارت‌ها رایج نیست و اغلب تنها به‌صورت کارت پرداخت برخط استفاده می‌شود. علاوه‌بر اینکه در عمل نیز، سامانه‌های کشف تقلب فعالی در کشور وجود ندارند و حتی نسبت به سایر کشورها، تحقیقات در این زمینه بسیار اندک است (وثوق، تقوی‌فرد و البرزی، ۱۳۹۳). با در نظر گرفتن وضعیت حاکم بر بزرگ‌داده، به‌ویژه امکان‌ناپذیری اجرای بسیاری از الگوریتم‌ها روی آن، تحقیقات صورت‌گرفته در کشور و سایر نقاط دنیا در این زمینه، بسیار محدود می‌شوند. یادآوری می‌شود که اصولاً داده‌ها زمانی بزرگ نامیده می‌شوند که سرعت تغییر، حجم و گوناگونی آنها بسیار بیشتر از توان سیستم فناوری اطلاعات موجود برای بازیابی، ذخیره، تحلیل و پردازش باشد (لوشن، ۲۰۱۳). کارت‌های بانکی،

یکی از اهداف مناسب برای متقلبان شناخته می‌شوند؛ زیرا چنانچه حمله‌کننده موفق شود، در زمان بسیار کوتاهی می‌تواند مقدار شایان توجهی پول برداشت کند، در حالیکه اغلب این برداشت در روزهای بعد شناسایی می‌شود (زری‌پور و شمس‌المعالی، ۲۰۱۵). همان‌گونه که گفته شد هدف مهم، شناسایی سریع تقلب و توقف آن در کوتاه‌ترین فاصلهٔ زمانی ممکن پس از رخدادن است (نصیری و مینایی، ۱۳۹۱)؛ به این معنا که بتوان بی‌درنگ آزمایش داده‌های تراکنش را انجام داد و رفتارهای مبهم کاربر را قبل از تکمیل تراکنش کشف کرد (حلوایی و اکبری، ۲۰۱۴). این انتقال پردازش داده از بعد به قبل از ذخیره‌سازی آن به شکل شایان توجهی زمان مقدور را برای ارزیابی تقاضاهای جدید از سیستم و به‌کارگیری تصمیم دقیقی برای کشف تقلب، کاهش می‌دهد (حلوایی و اکبری، ۲۰۱۴).

شایان ذکر است که روش‌های کشف تقلب آماری به دو زیرشاخهٔ با سرپرستی و بدون سرپرستی تقسیم می‌شوند (زری‌پور و شمس‌المعالی، ۲۰۱۵). یکی از دغدغه‌های پژوهشگران انتخاب رویکرد مناسب‌تر از میان این دو است (دال–پوزولو، کائلن، لوبرین، واترشوت و بونتمپی، ۲۰۱۴). از یک‌سو الگوریتم‌های با سرپرستی به برچسب‌گذاری تراکنش‌های قبلی نیاز دارند و معمولاً تنها به الگوهای تقلبی‌ای که در گذشته رخ داده است، محدود می‌شوند (دال–پوزولو و همکاران، ۲۰۱۴). از سوی دیگر، روش‌های غیرسرپرستی از طبقهٔ تراکنش‌ها استفاده نمی‌کنند و می‌توانند رفتارهای تقلب‌آمیز جدید را نیز شناسایی کنند. به‌علاوه، درخصوص هر نوع روشی که از رویکردهای با سرپرستی استفاده می‌کند، انتقادهایی وارد است: ۱. هزینهٔ محاسباتی زیادی دارند؛ ۲. زمان لازم برای برچسب‌زدن به مشاهدات جدید زیاد است؛ ۳. انحراف‌های ناشی از انتخاب نمونه می‌تواند سبب بروز خطا در برچسب‌های داده‌های آموزش شود (فوآ و همکاران، ۲۰۰۵).

با در نظر گرفتن این موارد، مسئلهٔ تحقیق حاضر، شناسایی سریع و درستی قابل قبول تقلب در تراکنش کارت‌های بانکی با استفاده از رویکرد غیرسرپرستی شبکهٔ عصبی نقشه‌های خودسازمانده کوهونن در محیط بزرگ‌داده است. مطابق بررسی‌های صورت‌گرفته، مدل پیشنهادشدهٔ این پژوهش با استفاده از شبکهٔ عصبی کوهونن و پردازش موازی در کشف ناهنجاری تراکنش کارت‌های پرداخت، نوعی نوآوری است و نمونهٔ مشابهی در پژوهش‌های قبلی نداشته است. یکی از رویکردهای استفاده از روش باسرپرستی برای تشکیل الگوی دارندهٔ کارت قانونی و متقلب، براساس یادگیری تراکنش‌های تاریخی و توسعهٔ مجموعه‌ای از قواعد است. در رویکرد غیرسرپرستی، بیشتر کشف الگو دنبال می‌شود؛ به این معنا که اگر تراکنشی به الگوی

دارندهٔ کارت قانونی مرتبط نباشد یا شبیه به الگوهای تقلب باشد، با عنوان مشکوک به تقلب طبقه‌بندی می‌شود (زاسلاوسکی و استریژاک، ۲۰۰۶).

پیشینهٔ پژوهش

پیشینهٔ نظری

به‌طور ساده، تقلب در کارت سوءاستفاده از آن بدون مجوز صاحب یا صادرکننده تعریف می‌شود (تریپاتی و راگا، ۲۰۱۳). چنانچه در محیط رقابتی تقلب شایع شود و سیستم پیش‌گیرنده و محدودکننده‌ای وجود نداشته باشد، می‌تواند به مسئلهٔ تهدیدکنندهٔ حیاتی برای کسب‌وکار و سازمان تبدیل شود (فوآ و همکاران، ۲۰۰۵). از سوی دیگر، با رشد سریع تعداد کارت‌های بانکی توزیع‌شده، طبیعی است که فعالیت‌های مجرمانه در این حوزه نیز افزایش یابد (زری‌پور، سیجا، علم و افشار، ۲۰۱۲). برخی از انواع شناخته‌شدهٔ تقلب در کارت‌های بانکی که معمولاً به‌صورت استفاده از شبکهٔ ارتباطی یا با حضور فیزیکی کارت صورت می‌پذیرند، عبارت‌اند از: تقلب‌های فروشنده، تقلب‌های اینترنتی، کارت گم‌شده یا ربوده‌شده، در اختیار گرفتن حساب، استفاده‌نکردن از کارت، دریافت‌نکردن کارت، جست‌وجو در سطل زباله، کارت‌های جعلی، سرقت پستی، افشای اطلاعات در محل کار یا منزل، شبکه‌های اجتماعی، تقلب ورشکستگی، تقلب دستگاه‌های خودپرداز و غصب‌کردن (وثوق و همکاران، ۱۳۹۳؛ نصیری و مینایی، ۱۳۹۱؛ زاسلاوسکی و استریژاک، ۲۰۰۶؛ سریواستاوا، کوندو، سورال و ماجومدار، ۲۰۰۸). هدف از شناسایی تقلب، متوقف‌کردن آن در کوتاه‌ترین فاصلهٔ زمانی ممکن پس از رویداد است (وثوق و همکاران، ۱۳۹۳). فارغ از اینکه سیستم کشف تقلب مد نظر، به‌صورت دستی یا سیستمی است، اصولاً باید از ویژگی‌های زیر برخوردار باشد:

- باید کاملاً درست تقلب را تشخیص دهد، یعنی نباید تراکنش اصلی با تراکنش تقلب‌آمیز اشتباه شود (درصد اعلان اشتباه آن پایین باشد)، در غیر این‌صورت ممکن است کاربران نسبت به هشدارهای سیستم بی‌تفاوت شوند (اکسلسون، ۲۰۰۰)؛
- باید در کمترین زمان ممکن تقلب را تشخیص دهد (نصیری و مینایی، ۱۳۹۱).

با وجود این، در زمینهٔ کشف تقلب چالش‌هایی وجود دارد که به‌طور خلاصه عبارت‌اند از:

- در دسترس‌نبودن مجموعه داده‌های واقعی: یکی از محدودیت‌های مهم در این زمینه، نداشتن مجموعه داده‌های واقعی برای آزمون به‌دلایل امنیتی و کسب‌وکار است (زری‌پور و شمس‌المعالی، ۲۰۱۵؛ کیبی و چونهوآ، ۲۰۱۱؛ گای و همکاران، ۲۰۱۱)؛

- مجموعه داده‌های نامتقارن: مجموعه داده‌های تراکنش کارت به‌شدت چولگی دارند؛ یعنی نسبت وجود تراکنش تقلب به تراکنش قانونی بسیار پایین است (کیبی و چونهوآ، ۲۰۱۱). معمولاً در وضعیت واقعی ۹۸ درصد از داده‌ها قانونی است و تنها ۲ درصد از آنها متقلبانه‌اند (زری‌پور و شمس‌المعالی، ۲۰۱۵؛ دال-پوزولو و همکاران، ۲۰۱۴؛ ویترو، هند، جوزاک، و وستن، ۲۰۰۸)؛
- اندازهٔ مجموعه داده‌ها: هر روز میلیون‌ها تراکنش کارت بانکی انجام می‌شود (چان و همکاران، ۱۹۹۹) و تحلیل چنین مقدار بی‌شماری از اطلاعات به روش‌های بسیار مؤثر و مقیاس‌پذیر محاسباتی نیاز دارد (زری‌پور و شمس‌المعالی، ۲۰۱۵)؛
- رفتار پویای متقلب: متقلبان رفتار پویایی دارند، یعنی در طی زمان رفتار خود را در مقابل سیستم تغییر می‌دهند. بنابراین با سپری شدن زمان، تقلب‌ها نیز پیچیده‌تر می‌شوند (زری‌پور و شمس‌المعالی، ۲۰۱۵؛ کوریا بنسن، اوآوا، استوجانویچ، و اوترستن، ۲۰۱۶)؛ کمابیش همهٔ تراکنش‌های متقلبانه ظاهر قانونی دارند و چنانچه هر یک از آنها جداگانه بررسی شود، هیچ نشانهٔ مشکوکی از آنها دیده نمی‌شود (وثوق و همکاران، ۱۳۹۳).

پیشینهٔ تجربی

در این بخش به پیشینهٔ عملی مطالعات صورت‌گرفته درخصوص کشف تقلب کارت‌های بانکی که معمولاً به‌عنوان کارت‌های اعتباری می‌شناسیم، پرداخته می‌شود. برای کشف تقلب در کارت‌های بانکی، از روش‌ها و الگوریتم‌های متعددی استفاده می‌شود که در زیر به بعضی از آنها اشاره شده است:

- الگوریتم‌های همجوشی اطلاعات مانند تئوری گواه دمپستر و شیفر و یادگیری بیزی (رآج و پرتیا، ۲۰۱۱؛ برمودز، پرز، آیوسو، گومز و وازکوئز، ۲۰۰۸)؛
- مدل مخفی مارکوف (سریواستاوا و همکاران، ۲۰۰۸؛ بوساری و پاتیل، ۲۰۱۱)؛
- شبکه‌های عصبی (زری‌پور و همکاران، ۲۰۱۲؛ رآج و پرتیا، ۲۰۱۱؛ وثوق و همکاران، ۱۳۹۳؛ زاسلاوسکی و استریژاک، ۲۰۰۶؛ گنزالس و ولاسکوئز، ۲۰۱۳؛ اولزووسکی، ۲۰۱۴؛ پتیدار و شارما، ۲۰۱۱)؛
- الگوریتم ژنتیک (راماکالیانی و اومادوی، ۲۰۱۲؛ دومان و ازلیک، ۲۰۱۱)؛
- سیستم‌های ایمنی مصنوعی (حلوایی و اکبری، ۲۰۱۴).

سریواستاوا و همکارانش (۲۰۰۸) برای شناسایی تقلب در کارت‌های بانکی، از مدل مخفی مارکوف محدود با بهره‌مندی از رویکرد کشف ناهنجاری رفتاری استفاده کردند. اپیلارد و بوگیلا (۲۰۱۶) با استفاده از مدل مخفی مارکوف در داده‌های به‌دست‌آمده از پردازش تصویر دوربین‌های شهری، به‌دنبال کشف ناهنجاری بودند. وثوق و همکارانش (۱۳۹۳) برای تشخیص سریع تقلب در تراکنش‌های بانکی از مدل شبکه‌های عصبی مصنوعی چندلایهٔ جلوسو بهره بردند. آنها از داده‌های واقعی استفاده کردند، اما به‌دلیل نداشتن برچسب، تراکنش‌های مشکوک را به‌کمک دانش خبرگان و ادبیات موضوع شبیه‌سازی کردند.

اولزووسکی (۲۰۱۴) به‌کمک مدل شبکهٔ عصبی نقشه‌های خودسازمان‌ده، چارچوبی به‌منظور کشف تقلب ارائه داد. او با استفاده از بصری‌سازی رفتار کاربر و یک ماتریس یو[1]، آستانهٔ سنجش رفتار ناهنجار را تعیین کرد. گنزالس و ولاسکوئز (۲۰۱۳) در کار خود، برای مشخص‌کردن فرارهای مالیاتی از نوعی نقشهٔ خودسازمانده برای ایجاد مدل کلی طبقه‌بندی استفاده کردند. هوانگ، سای و یو (۲۰۱۴) در پژوهش خود از یک کوهونن سلسله‌مراتبی رشدیابنده دوگانه استفاده کردند و آن را GHSOM نامیدند. این مدل برای کشف ناهنجاری در گزارش‌های مالی به‌کار رفته است. هلمن، ترسپ، و سیمولا (۱۹۹۹) بر اساس مدل شبکهٔ عصبی نقشهٔ خودسازمان دهنده، نوعی سیستم کشف تقلب ارائه دادند که در آن مدل پس از آموزش با استفاده از مدل‌های احتمالی پروفایل کاربر، تقلب‌های احتمالی را براساس کشف ناهنجاری خوشه‌بندی‌محور، شناسایی می‌کند. زاسلاوسکی و استریژاک (۲۰۰۶) با استفاده از مدل کوهونن، برای صاحب کارت پروفایل رفتار عادی ساختند. آنها ابتدا با استفاده از تراکنش‌های قبلی و شبکهٔ کوهونن، ماتریس وزن‌ها را به‌عنوان پروفایل ذخیره کردند، سپس با اندازه‌گیری فاصلهٔ تراکنش ورودی با پروفایل به‌دست‌آمده، میزان عادی یا ناهنجار بودن آن را بررسی کردند.

کوا و سریگانش (۲۰۰۸) با استفاده از شبکه‌های عصبی مدل کوهونن، سیستم خود را به سه لایه تقسیم کردند. مدل آنها ترکیبی و نیمه‌سرپرستی است و در آن از یک شبکهٔ عصبی کوهونن به همراه شبکهٔ عصبی پرسپترون استفاده شده است. درخصوص کشف ناهنجاری در بزرگ‌داده نیز می‌توان به کار حلوایی و اکبری (۲۰۱۴) اشاره کرد که با استفاده از سیستم‌های ایمنی مصنوعی و مدل نگاشت کاهش در محیط پردازش ابر، نوعی مدل کشف تقلب با سرپرستی ارائه دادند. هوانگ، ژو، یانگ و فنگ (۲۰۱۶) برای کشف ناهنجاری براساس رویکردهای چگالی‌محور، روش جست‌وجویی به‌نام جست‌وجوی همسایگی محلی ارائه دادند. به، وانگ، زین

1. U-Matrix

و وانگ (۲۰۱۶) نیز بر کشف ارزش‌های ناهنجار در محیط‌های بزرگ‌دادهٔ توزیع‌شده تمرکز کردند. روش استفاده‌شدهٔ آنها کشف ناهنجاری چگالی‌محور است.

با بررسی پیشینهٔ عملی مشاهده می‌شود، هیچ‌یک از مدل‌های بدون سرپرستی شبکهٔ عصبی کوهونن، به‌صورت ترکیب با نگاشت کاهش برای غلبه بر الزامات بزرگ‌داده استفاده نکرده‌اند.

مدل مفهومی

همان‌گونه که پیش از این بیان شد، مسئلهٔ تحقیق شناسایی سریع و با درستی قابل قبول تقلب در تراکنش کارت‌های بانکی است.

در نظر بگیرید که در یکی از پایگاه‌های داده، مجموعه‌ای مانند $C = \{\vec{c}_1, \cdots, \vec{c}_{k_n}\}$ به‌عنوان مجموعه کارت‌های ذخیره‌شده در سیستم پرداخت و زیرمجموعه $\vec{c}_i = \{1, \cdots, c_s^i\}$، کارتی متشکل از s ویژگی است. در ضمن، مجموعهٔ بزرگ $X = \{|X|^1, \cdots, |X|^{k_s}\}$، کل تراکنش‌های موجود در سیستم و $|X|^i = \{\vec{X}_1^i, \cdots, \vec{X}_n^i\}$، رکوردهای مربوط به کارت شمارهٔ i تا زمان t_n است. هر یک از این رکوردها، بردارهایی مانند $\vec{X}_j = \{x_j^1, \cdots, x_j^m\}$، هستند که از m ویژگی عددی یا اسمی تشکیل شده‌اند. $\vec{X}_{n+1}, \cdots, \vec{X}_{n+k}$ نیز تراکنش‌های پس از زمان t_n هستند. حال برای کارتی مانند i، مسئلهٔ یافتن قانونی یا تقلب‌آمیز بودن تراکنش $\vec{X}_{n+1}$ در کمترین زمان ممکن است. مجموعهٔ X از تراکنش‌های سیستم پرداخت به دو زیرمجموعهٔ جمع‌ناپذیر تقسیم قانونی $X^l \subseteq X$، و تقلب‌آمیز $X^f \subseteq X$ به‌طوری‌که $X^f \cap X^l = \emptyset$ تقسیم می‌شود. اگر فرض بر این قرار گیرد که تصویر عددی (نقاط موجود در برخی از فضاهای چندبعدی) از تراکنش‌های قانونی و تقلب‌آمیز به زیرفضاهای متفاوتی تعلق دارند؛ ممکن است دربارهٔ نگاشت تراکنش جدیدی مانند $\vec{X}_{n+1}$ به هر یک از زیرفضاها، تصمیم‌گیری شود. فضای تصمیم برای طبقه‌بندی تراکنش جدید از رابطهٔ ۱ به‌دست می‌آید.

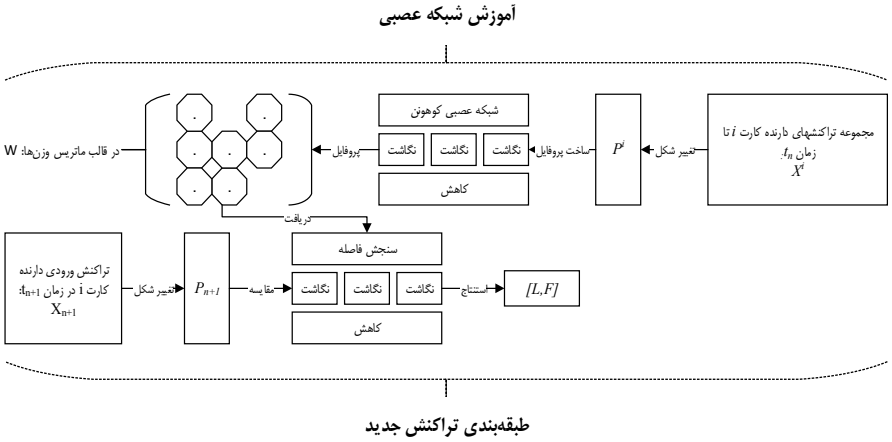
$$\Theta = \{\theta_f, \theta_l\} \qquad \text{رابطهٔ ۱)}$$

در این رابطه، θ_f و θ_l به‌ترتیب تقلب‌آمیز یا قانونی‌بودن تراکنش هستند. بنابراین، عضویت تراکنش به هر یک از زیرفضاهای یادشده براساس میزان مشابهت $\vec{X}_{n+1}$، به زیرمجموعهٔ $X^l \subseteq X_n$؛ و از طریق آموزش یک پروفایل از رفتار هنجار دارندهٔ کارت، سنجیده می‌شود. بنابراین، با توجه به هدف تحقیق، مدل ریاضی به‌صورت رابطهٔ ۲ است که در آن E خطای طبقه‌بندی، t_d زمان یافتن پاسخ و $g(X_{n+1})$ تابع طبقه‌بندی است.

رابطهٔ ۲)

$$min\, z = \langle t_d, E \rangle$$
$$st: g(X_{n+1}) \in \Theta$$

شکل ۱ مدل مفهومی پژوهش را با در نظر گرفتن اهداف آن، یعنی دقت و سرعت در طبقه‌بندی نشان می‌دهد.



شکل ۱. مدل مفهومی پیشنهادی کشف تقلب در تحقیق حاضر

روش‌شناسی پژوهش

در این بخش به تعیین ویژگی‌های لازم برای تشکیل پروفایل رفتار دارندهٔ کارت، گردآوری و پیش‌پردازش داده‌ها و مدل‌سازی حل تابع هدف مسئلهٔ تحقیق پرداخته خواهد شد.

ویژگی‌های رفتاری

کشف تقلب با توجه به تحلیل عادت استفاده از کارت، راه‌حل امیدوارکننده‌ای است (سریواستاوا و همکاران، ۲۰۰۸)؛ به این معنا که افراد معمولاً تمایل دارند بعضی از ویژگی‌های خاص رفتاری را به نمایش بگذارند. بنابراین، هر دارندهٔ کارت می‌تواند به‌وسیلهٔ مجموعه الگوهایی که انحراف از آنها رفتار مشکوک بالقوه‌ای است، به نمایش درآیند (زاسلاوسکی و استریژاک، ۲۰۰۶). هنگام ساخت مدل کشف تقلب، مجموعهٔ اولیهٔ ویژگی‌ها شامل اطلاعات مربوط به تراکنش تکی است (کوریا بنسن و همکاران، ۲۰۱۶). مدل‌های بسیاری از این ویژگی‌های خام استفاده کرده‌اند (مینگیشی و نیمی، ۲۰۱۱؛ سریواستاوا و همکاران، ۲۰۰۸)، اما استفاده از این داده‌های خام به

بیرون‌کشیدن رفتار مالک منجر نخواهد شد (کوریا بنسن و همکاران، ۲۰۱۶). ویترو و همکارانش (۲۰۰۸) برای به تصویرکشیدن عادت رفتاری صاحب کارت، راهبرد استفاده از ویژگی‌های تجمیع‌شده را ارائه دادند. به عقیدهٔ کوریا بنسن و همکارانش (۲۰۱۶)، راه بهتر، تکمیل رویکرد ویترو و استفاده از ویژگی‌های تجمیعی تراکنشی است. این ویژگی‌های تجمیعی شامل گروه‌بندی تراکنش‌های انجام‌شده در آخرین دورهٔ زمانی به تفکیک اشاره‌شده در بالا می‌شوند. این رویکرد در پژوهش‌های متعددی استفاده شده است (زاسلاوسکی و استریژاک، ۲۰۰۶؛ جها، گویلن، و وستلند، ۲۰۱۲؛ دال–پوزولو و همکاران، ۲۰۱۴؛ کوریا بنسن و همکاران، ۲۰۱۶). بر همین اساس، در این تحقیق پس از بررسی ویژگی‌های یادشده توسط خبرگان، ویژگی‌های ترکیبی مانند، جمع ارزش پولی تراکنش، تعداد تراکنش در t ساعت گذشته و... و ارزش پولی تراکنش و میانگین ساعت‌های استفاده در قالب هشت ویژگی برای استخراج الگوی رفتاری دارندهٔ کارت در نظر گرفته شد.

گردآوری و پیش‌پردازش داده‌ها

برای این تحقیق از داده‌های واقعی یکی از بانک‌های فعال کشور که از کارت‌های الکترونیکی استفاده می‌کند، با در نظر گرفتن شرایط محرمانگی، استفاده شد. تراکنش‌های ۱۵۷۹۷ کارت در یک بازهٔ چهارماهه به‌صورت تصادفی انتخاب شد. در گام نخست تعداد تراکنش‌ها ۷۲۲۹۸۸ بود. با توجه به اثربخش‌نبودن مدل با تعداد تراکنش اندک برای دارندهٔ کارت (زاسلاوسکی و استریژاک، ۲۰۰۶)، کارت‌هایی که در این بازه کمتر از ۵۰ تراکنش داشتند، حذف شدند. از سوی دیگر، با بررسی اسناد بانک یادشده، تعدادی تراکنش تقلب‌آمیز کشف شد. آنها تراکنش‌هایی بودند که مشتریان به‌عنوان غیرقابل قبول و فهرست سیاه معرفی کردند بودند. با توجه به تعداد اندک آنها و پس از بررسی الگوهایشان و به‌کارگیری نظر خبرگان، به‌منظور اعتبارسنجی مدل برای رسیدن به مقدار ۲ درصد تراکنش تقلب‌آمیز، نویزی با توزیع نزدیک به نظر خبرگان و الگوی تراکنش‌های تقلب‌آمیز استخراج شد و به آن تعدادی تراکنش قانونی داده شد و آنها را برچسب‌دار کرد. در گام بعدی تراکنش‌های قانونی‌ای که ارزش‌های پرت داشتند از مجموعهٔ داده حذف شدند. پس از حذف این تراکنش‌ها، درنهایت تعداد کارت‌ها به ۸۹۰۰ و تراکنش‌ها به ۵۸۰۵۰۸ رسید. ۹۰ درصد از تراکنش‌ها برای ساخت پروفایل و ۱۰ درصد برای آزمون مدل استفاده شدند. در مجموع، در داده‌های آزمون ۱۱۲۲ تراکنش تقلب‌آمیز وجود داشت. از آنجا که در این تحقیق از ویژگی‌های خام تراکنش استفاده نمی‌شود و به‌جز مقدار پولی تراکنش، باقی ویژگی‌ها به‌صورت ترکیبی در یک بازهٔ زمانی هستند، از رابطهٔ ۳ به‌منظور تبدیل‌کنندهٔ تراکنش‌های خام به شکل ترکیبی تراکنش استفاده می‌شود.

$$Tr(X, Du) = P: R^s \rightarrow R^8$$ رابطهٔ ۳)

در رابطهٔ ۳، X مجموعه تراکنش‌های یک کارت، و Du بازهٔ زمانی مد نظر برای تجمیع و P صورت تغییریافتهٔ تراکنش‌هاست. به‌علاوه، محاسبهٔ ویژگی میانگین ساعت‌های تراکنش برای یک دوره، با توجه به کارایی‌نبودن میانگین با پیشنهاد کوریا بنسن و همکارانش از توزیع ون میزز[1] (بیشاپ، ۲۰۰۶) استفاده شد (کوریا بنسن و همکاران، ۲۰۱۶). درنهایت برای نرمال‌سازی نیز از مبدل دوقطبی استفاده شده است.

ساخت مدل

در این بخش برای دستیابی به هدف تحقیق به ساخت مدل پرداخته می‌شود. همان‌گونه که بیان شد، رویکرد تحقیق استفاده از مدل‌های بدون‌سرپرستی است، بنابراین از شبکهٔ عصبی کوهونن برای ایجاد پروفایل دارندهٔ کارت و سنجش میزان انحراف تراکنش جدید از پروفایل ایجادشده و از رویکرد نگاشت کاهش برای موازی‌سازی اجرای این مدل استفاده شده است.

نقشه‌های خودسازمانده[2]، شبکه‌های عصبی‌ای هستند که می‌توانند فضای ورودی چندبعدی را به فضای خروجی دوبعدی نگاشت کنند (کوهونن، ۱۹۹۰). مدل کوهونن دو فاز دارد؛ در فاز آموزش، نقشه‌ها ساخته می‌شوند و در فاز پیش‌بینی، یک بردار ورودی به نزدیک‌ترین نقشه نگاشت می‌شود (کوهونن، ۱۹۹۰). در این شبکه با واردشدن بردار، تنها بخشی از شبکه به‌روز می‌شود (رجاس، ۱۹۹۶). فرایند خودسازماندهی براساس آموزش رقابتی است و شامل تنظیم ماتریس وزن‌ها یا W^{i} تشکیل‌شده از k (تعداد نرون‌ها) بردار ستونی m (تعداد ویژگی‌ها) در یک فرایند تکراری، با توجه به بردار ورودی می‌شود (کوهونن، ۱۹۹۰). در این تحقیق، ماتریس وزن‌ها پس از آموزش شبکه با استفاده از قاعدهٔ یادگیری کوهونن (کوهونن ۱۹۹۰) ایجاد می‌شود.

$$\overrightarrow{\mathrm{w}}_l^{t+1} = \overrightarrow{\mathrm{w}}_l^{t+1} + h(t).\eta(t)(\overrightarrow{\mathrm{x}}_l^t - \overrightarrow{\mathrm{w}}_l^t)$$ رابطهٔ ۴)

که در آن، t زمان یا تکرار، $\overrightarrow{w}_l$ نماد lامین $(\mathrm{l} = ۱, \dots, \mathrm{k})$ بردار m عضوی از ماتریس وزن‌ها، h تابع همسایگی (تابعی از زمان و R یا شعاع همسایگی)، η ضریب یادگیری و $\vec{x}_l$ بردار ورودی m عضوی مرتبط با نرون lام است. جدول ۱ به‌صورت خلاصه الگوریتم یادگیری کوهونن را نشان می‌دهد.

---

1. Von Mises distribution
2. Self organizing maps (SOM)

جدول ۱. الگوریتم یادگیری کوهونن

آغاز الگوریتم:

۱. وزن‌دهی تصادفی ماتریس وزن‌ها (W)، مقداردهی ضریب یادگیری اولیه η.، شدت همسایگی اولیه σ.، شعاع همسایگی R و تابع همسایگی h.

۲. برای همهٔ بردارهای ورودی ($\vec{x}_i$) به ازای i=۱...n (n، تعداد مشاهدات است):

- سنجش فاصلهٔ بردار انتخاب‌شده با همهٔ نرون‌ها و یافتن بهترین نرون برازنده[1]

رابطهٔ ۵) $$BFN = Argmin\left(d\left(\vec{x}_i, \vec{w}_j\right); j = ۱ \cdots k\right)$$

در رابطهٔ ۵، d(.,.) معیار فاصله است.

- با در نظر گرفتن شعاع همسایگی، به‌روزرسانی وزن‌های بهترین نرون برازنده و همسایه‌هایش با استفاده از رابطهٔ ۱.

۳. به‌روزرسانی، ضریب یادگیری و شدت همسایگی

۴. سنجش شرط پایان حلقه (معمولاً تعداد تکرار) در غیر این‌صورت رفتن به گام ۲

پایان الگوریتم

منبع: کتو و دشپاند (۲۰۱۵)

در فاز پیش‌بینی، بردار ورودی جدید ($\vec{x}_{n+۱}$) با همهٔ نرون‌ها سنجیده می‌شود و نزدیک‌ترین نرون، به‌عنوان برندهٔ این بردار ورودی، شناخته خواهد شد.

رابطهٔ ۶) $$Cluster = Argmin\left(d\left(\vec{x}_{n+1}, \vec{w}_j\right); j = 1 \cdots k\right)$$

قبل از معرفی روش پیشنهادی، مختصری دربارهٔ نگاشت کاهش که اساسی برای پردازش توزیعی بزرگ‌داده‌هاست، توضیح داده می‌شود. نگاشت کاهش نوعی مدل توسعهٔ سیستم برای محاسبات موازی و توزیع شده است که حجم‌های عظیم داده‌ها را مدیریت می‌کند. این مدل دو عملگر دارد که به‌صورت مجموعه یا زوج‌های کلید/ ارزش، نتایج را بازمی‌گردانند (دین و قماوات، ۲۰۰۴). عملگر نگاشت اجرای بخشی از تحلیل را برعهده دارد و محاسبه یا تحلیل‌ها را در مجموعه‌های از زوج‌های کلید/ ارزش میانی وارد می‌کند (دین و قماوات، ۲۰۰۴). کاهش این عملگر، مجموعه زوج‌های تولیدشده توسط نگاشت‌ها را به‌عنوان ورودی دریافت‌کرده و برای دستیابی به نتیجهٔ نهایی تحلیل، پس از پردازش آنها، این مجموعه‌ها را با کلید نهایی مرتبط می‌کند. با استفاده از این عملگرها، هر برنامهٔ نگاشت کاهش می‌تواند مجموعه‌ای متوالی از وظایف موازی را روی حجم عظیمی از داده‌ها انجام دهد (لسکوک، راجارمان، و اولم، ۲۰۱۴). ترکیب دو ویژگی استقلال داده‌ها و استقلال محاسباتی، این دست‌آورد را به ارمغان می‌آورد که بتوان وظایف را به‌صورت توزیع‌شده و موازی پیاده‌سازی کرد. این قابلیت به توسعه‌دهنده اجازه

1. Best fitted neuron (BFN)

می‌دهد که از مقیاس‌پذیری نتایج پردازش موازی در مقیاس عظیم، برای افزایش سرعت پردازش و عملکرد بهره‌مند شود (لسکوک و همکاران، ۲۰۱۴).

برای ایجاد پروفایل، دارندهٔ کارت i با استفاده از مدل کوهونن، X^{i} نشان‌دهندهٔ رکوردهای مربوط به کارت i تا زمان t_{n} است که از طریق رابطهٔ ۳ و نرمال‌سازی به مجموعهٔ تجمیعی P^i تبدیل می‌شود. برای تولید ماتریس k برداری، W^{i} که وزن‌های آموزش‌دیدهٔ نرون‌های به‌دست‌آمده از آموزش شبکه براساس ورودی‌های ماتریس P^i است؛ از الگوریتم پیشنهادی مندرج در جدول ۲، استفاده می‌شود.

جدول ۲. الگوریتم پیشنهادی برای آموزش شبکهٔ عصبی کوهونن براساس نگاشت کاهش

آغاز الگوریتم:

۱. وزن‌دهی تصادفی ماتریس وزن‌ها (W)، مقداردهی ضریب یادگیری اولیه η، نرخ شدت همسایگی اولیه σ، شعاع همسایگی R و تابع همسایگی h.

۲. برای هر بردار ورودی p_j^i (i شمارهٔ کارت، و j=۱...n تعداد بردارهای ورودی)، فرایند نگاشت به تعداد M تا رسیدن به آخرین نرون به‌صورت موازی اجرا می‌شود ($M \leq k$، k تعداد نرون‌هاست):

- نگاشت به‌صورت زیر تعریف می‌شود:

رابطهٔ ۷) $$map(p_j^i, w_h^i) =< h, d(p_j^i, w_h^i) >$$

- رابطهٔ ، فاصلهٔ jامین ورودی را با hامین نرون می‌سنجد و مقدار فاصله را در قالب زوج کلید و ارزش <شمارهٔ نرون، میزان فاصله> باز می‌گرداند.
- در گام کاهش، پس از به پایان رسیدن همهٔ نگاشت‌ها در قالب تک‌فرایند نخست، بهترین نرون برازنده با دادن فهرستی از کلید، ارزش‌های مجموعهٔ نگاشت به‌صورت رابطهٔ محاسبه می‌شود:

رابطهٔ ۸) $$BFN(list < h, d(p_j^i, w_h^i) >) = Key \min_{h=1\cdots k}(v_h)$$

- سپس در همین گام کاهش با در نظر گرفتن شعاع همسایگی، به‌روزرسانی وزن‌های بهترین نرون برازنده و همسایه‌هایش با استفاده از رابطهٔ ۴ صورت می‌پذیرد.

۳. به‌روزرسانی، ضریب یادگیری و شدت همسایگی

۴. سنجش شرط پایان حلقه (معمولاً تعداد تکرار) در غیر این‌صورت رفتن به گام ۲

پایان الگوریتم

شایان ذکر است، تابع فاصله همان تابع فاصلهٔ اقلیدسی اشاره‌شده در رابطهٔ ۹ است.

رابطهٔ ۹) $$d(\vec{P}, \vec{W}) = \sqrt{\sum_{i=1}^{m}(p_i - w_i)^2}$$

پس از طی فرایند یادشده برای همهٔ کارت‌ها و براساس مجموعهٔ آموزش که تشکیل‌شده از صورت تغییریافتهٔ تجمیعی و نرمال‌شدهٔ مجموعهٔ آموزش است؛ نوبت به سازوکاری برای سنجش احتمال ناهنجاری و متعاقب آن، تقلب تراکنش ورودی براساس ماتریس W^{i} به‌دست‌آمده است.

در ادامه، نخست با توجه به اینکه فاصلهٔ معیاری پیوسته و طبقه‌بندی معمولاً دوقطبی است، راه‌حل‌های موجود درخصوص تراکنش ناهنجار بررسی می‌شود، دوم روش پیشنهادی این پژوهش برای تشخیص تراکنش ناهنجار ارائه‌شده و الگوریتمی برای سنجش موازی فاصلهٔ تراکنش ورودی با پروفایل آموزش‌دیده، معرفی خواهد شد.

یکی از رویکردهایی که برای محاسبهٔ مقدار مشابهت میان $\vec{X}_{n+1}$ با W^{i} به‌کار می‌رود، یافتن نزدیک‌ترین نرون و مقایسهٔ فاصلهٔ تراکنش ورودی با آن نرون با آستانهٔ یکسان است. محققان معتقدند هرچه فاصله بیشتر باشد، احتمال تقلب نیز بیشتر خواهد بود (زاسلاوسکی و استریژاک، ۲۰۰۶). در رویکرد دوم، نرونی که بیشترین عدد را در ماتریس یو[1] به خود تخصیص می‌دهد، به‌عنوان مرز آستانه در نظر گرفته می‌شود. ماتریس یو، شکلی از نمایش در توری شبکهٔ کوهونن است که همهٔ نرون‌ها وارد آن می‌شوند و پس از سنجیدن تفاوت آنها با همسایه‌هایشان، نرون‌ها از کمترین بی‌تشابهی با همسایه‌ها تا بیشترین مرتب می‌شوند. بنابراین، در این ماتریس نوعی توالی از ارزش‌های بالا با پراکندگی زیاد وجود دارد. با توجه به این ماتریس، آستانه به‌صورت رابطهٔ ۱۰ محاسبه می‌شود. در رابطهٔ ۱۰، τ ارزش آستانه، C مرکز کلی توری شبکه، v_{max} نرونی بیشترین ارزش در ماتریس یو را دارد (نرونی که درصد مشابه‌نبودن آن با همسایه‌هایش بیشتر است) و $\mathrm{d}(.,.)$ معیار فاصله است (اولزووسکی، ۲۰۱۴).

$$\tau = d(C, v_{max}) \qquad \text{رابطهٔ ۱۰)}$$

در این رویکرد برای یافتن تراکنش ناهنجار، تابع منطقی رابطهٔ ۱۱، برای همهٔ ورودی‌ها محاسبه می‌شود. در این رابطه، p_{n+1} تراکنش ورودی تغییریافتهٔ جدید، v_{n+1} نرون مربوط به تراکنش ورودی ذکرشده و φ تابع منطقی مقایسه است.

$$\varphi(p_{n+1}) = \begin{cases} true, d(v_{n+1}, C) > \tau \\ false, d(v_{n+1}, C) \leq \tau \end{cases} \qquad \text{رابطهٔ ۱۱)}$$

در رویکرد پیشنهادی، مقدار به‌دست‌آمده از رابطهٔ ۱۰ در نظر گرفته می‌شود، سپس به‌کمک رابطهٔ ۱۲، شاخص دیگری با عنوان میانگین فاصلهٔ نامتشابه‌ترین نرون، محاسبه می‌گردد.

$$\rho = mean(d_{j=1,\cdots,\pm R}(v_{max}, v_j) \qquad \text{رابطهٔ ۱۲)}$$

1. U-Matrix

حال فاصلهٔ تراکنش ورودی با نزدیک‌ترین نرون محاسبه می‌شود. رابطهٔ ۱۳ تابع منطقی برای بررسی قانونی یا تقلب‌آمیز بودن تراکنش را نشان می‌دهد.

$$\acute{\varphi}(p_{n+1}) = \begin{cases} true, d(v_{n+1}, p_{n+1}) > \tau + \rho \\ false, d(v_{n+1}, p_{n+1}) \leq \tau + \rho \end{cases}$$ رابطهٔ ۱۳)

برای سنجش طبقه‌بندی موازی تراکنش ورودی براساس پروفایل، پس از تغییر شکل تراکنش، از الگوریتم پیشنهادی جدول ۳ استفاده می‌شود.

جدول ۳. الگوریتم پیشنهادی برای سنجش فاصلهٔ تراکنش ورودی با پروفایل دارندهٔ کارت

برای بردار ورودی P^i_{n+1} (i شمارهٔ کارت)، فرایند نگاشت به تعداد M تا رسیدن به آخرین نرون به‌صورت موازی اجرا می‌شود ($M \leq k$، k تعداد نرون‌هاست):

نگاشت به‌صورت زیر تعریف می‌شود:

$$map(P^i_{n+1}, w^i_h) =< h, d(P^i_{n+1}, w^i_h) >$$ رابطهٔ ۱۴)

رابطهٔ ۱۴، فاصلهٔ jامین ورودی را با hامین نرون می‌سنجد و مقدار فاصله را در قالب زوج کلید و ارزش <شمارهٔ نرون، میزان فاصله> بازمی‌گرداند.

در گام کاهش، پس از به پایان رسیدن همهٔ نگاشت‌ها در قالب تک‌فرایند؛ نخست فاصلهٔ تراکنش ورودی با نزدیک‌ترین نرون را یافته و همزمان فاصلهٔ نرون یادشده با مرکز توری شبکه، سنجیده می‌شود:

$$Reduce(list < h, d(P^i_{n+1}, w^i_h) >) = \langle \min_{h=1\cdots k}(v_h), d(v_{n+1}, C) \rangle$$ رابطهٔ ۱۵)

یافته‌های تحقیق

برای سنجش عملکرد الگوریتم این تحقیق در طبقه‌بندی، معیارهای طبقه‌بندی استفاده می‌شوند. جدول ۴ ماتریس درهم‌ریختگی طبقه‌بندی را به تصویر می‌کشد. بر همین اساس معیارهای جدول ۵ محاسبه می‌شوند.

جدول ٤. ماتریس درهم‌ریختگی الگوریتم‌های طبقه‌بندی

کل جامعه	شرایط درست	شرایط نادرست
پیش‌بینی مثبت	اعلان درست[1]	اعلان نادرست[2]
پیش‌بینی منفی	عدم اعلان نادرست[3]	عدم اعلان درست[4]

منبع: سوئت (۱۹۹۶)

1. True Positive (TP)
2. False Positive (FP)
3. False Negative (FN)
4. True Negative (TN)

جدول ۵. شاخص‌های مختلف سنجش عملکرد طبقه‌بندی

نام شاخص	فرمول
صحت یا درستی	$Acc = \frac{\sum TP + \sum TN}{Total\ Population}$
اعلان درست یا حساسیت یا فراخوان[1]	$TPR = \frac{\sum TP}{\sum P}$
عدم اعلان نادرست یا از دست دادن[2]	$FNR = \frac{\sum FN}{\sum P}$
اعلان نادرست یا مشاجره[3]	$FPR = \frac{\sum FP}{\sum N}$
عدم اعلان درست یا ویژگی[4]	$TNR = \frac{\sum TN}{\sum N}$
آمارهٔ F_1	$F_1 = \frac{2TP}{2TP + FP + FN}$
$g\text{-}mean_2$	$\sqrt{TPR \times TNR}$

منبع: پاورز (۲۰۱۱)

ابتدا براساس الگوریتم جدول ۱، شبکهٔ کوهونن را آموزش‌داده و داده‌های آزمون وارد آن می‌شوند. برای ارزش برش، از عدد ۱ تا ۱۰ درصد استفاده شد. جدول ۶ نتایج به‌دست‌آمده را در هر سطح برش به تصویر می‌کشد.

جدول ٦. نتیجهٔ آزمون براساس مقادیر برش (به درصد) و فاصله با نزدیک‌ترین نرون

شاخص	۱	۲	۳	٤	٥	٦	۷	۸	۹	۱۰
حساسیت	۲۵	۳۵	۴۱	۴۶	۶۰	۷۳	۷۴	۷۴	۷۷	۸۰
از دست دادن	۲۹	۳۶	۴۰	۵۳	۵۶	۶۱	۶۵	۶۹	۷۳	۷۷
درستی	۷۰	۶۴	۶۰	۵۲	۴۷	۴۰	۳۹	۳۳	۲۸	۲۶
F_1	۳	۴	۴	۳	۴	۵	۴	۴	۴	۴
$g\text{-}mean_2$	۴۳	۴۷	۵۰	۴۶	۵۱	۵۴	۵۱	۴۸	۴۵	۴۳

1. Recall
2. Miss
3. Fall out
4. Specificity

همان‌گونه که جدول ۶ نشان می‌دهد، استفاده از رویکرد مقدار برش ثابت برای همهٔ پروفایل‌ها و سنجش فاصله با نزدیک‌ترین نرون، طبقه‌بندی مناسبی انجام نمی‌دهد. با در نظر گرفتن اطلاعات این جدول، بیشترین حساسیت و از دست‌دادن، مربوط به مقدار برش ۱۰ درصد است. از سوی دیگر، بیشترین درستی برای برش ۱ درصد و بیشترین مقدار $g\text{-}mean_۲$ مختص به برش ۶ درصد است. در خصوص آمارهٔ $F_۱$ اساساً این روش در جایگاه مناسبی قرار ندارد. در گام دوم و سوم از روش استفاده‌شدهٔ اولزووسکی (۲۰۱۴) و رویکرد پیشنهادی برای طبقه‌بندی استفاده شد. جدول ۷ ماتریس درهم‌ریختگی براساس روش زاسلاوسکی و جدول ۸ این ماتریس را برای روش پیشنهادی نشان می‌دهد.

جدول ۹ نیز نتایج حاصل را برای هر دو روش به تصویر می‌کشد. نکتهٔ شایان توجه اینکه هر دو روش عملکرد چشمگیری را از خود نشان دادند. البته، روش پیشنهادی عملکرد مناسب‌تری دارد، به‌طوری‌که توانسته است به صحت ۹۴ درصد دست یابد. مهم‌ترین نکتهٔ موفقیت این دو روش را می‌توان در استفاده تطبیقی ارزش برش برای هر دارندهٔ کارت به‌صورت جداگانه دانست.

جدول ۷. ماتریس درهم‌ریختگی برای طبقه‌بندی براساس روش اولزووسکی

	شرایط درست	شرایط نادرست
پیش‌بینی مثبت	۹۰۱	۹۴۸۲
پیش‌بینی منفی	۲۲۱	۴۲۶۷۰

جدول ۸. ماتریس درهم‌ریختگی برای طبقه‌بندی با روش پیشنهادی

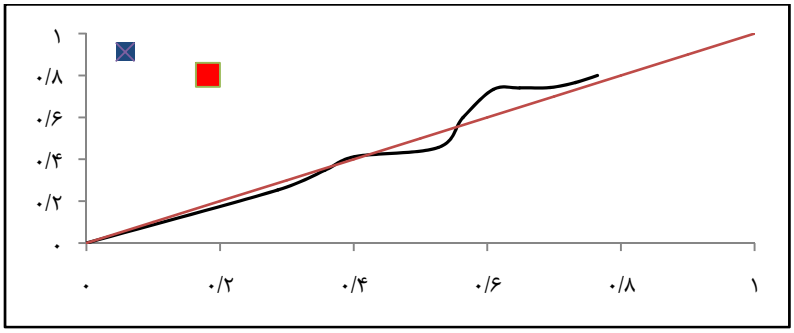
	شرایط درست	شرایط نادرست
پیش‌بینی مثبت	۱۰۲۴	۳۱۲۵
پیش‌بینی منفی	۹۸	۴۹۱۲۵

جدول ۹. نتیجهٔ آزمون شبکه براساس روش‌های اولزووسکی و پیشنهادی (ارقام به درصد)

روش	حساسیت	از دست دادن	درستی	$F_۱$	$g\text{-}mean_۲$
اولزووسکی	۸۰	۱۸	۸۲	۱۶	۸۱
پیشنهادی	۹۱	۶	۹۴	۴۰	۹۲

شکل ۲ نمودار ROC هر سه رویکرد را نشان می‌دهد. سطح زیر منحنی رویکرد اول برابر ۶۹ درصد است. به‌علاوه، دو نقطهٔ سمت چپ بالا، مربوط به رویکردهای تطبیقی است. در این شکل محور افقی نشان‌دهندهٔ مشاجره و محور عمودی حساسیت است.

پس از بررسی معیارهای طبقه‌بندی، نوبت به مقایسهٔ عملکرد الگوریتم تک‌فرایندی جدول ۱، با الگوریتم ترکیبی موازی دو سطحی جدول ۲ برای آموزش شبکه و جدول ۳ برای طبقه‌بندی تراکنش ورودی از نظر زمان اجرا می‌رسد. پارامتری که مقیاس‌پذیری مدل کوهونن را تحت تأثیر قرار می‌دهد، تعداد نرون‌های شبکه است. بنابراین، تعداد نرون‌ها از ۱۰ تا ۵۰ با گام‌های ۱۰تایی در نظر گرفته‌شده و زمان اجرا، بررسی شد. جدول ۱۰ نتایج را نشان می‌دهد. همان‌گونه که در این جدول ملاحظه می‌شود با بالارفتن تعداد نرون‌ها، زمان لازم برای ساخت پروفایل و طبقه‌بندی تراکنش جدید به‌صورت چشمگیری افزایش می‌یابد.



شکل ۲. نمودار ROC مربوط به هر سه روش اندازه‌گیری فاصله

جدول ۱۰. نتیجه آزمون الگوریتم تک‌فرایندی (اعداد براساس میلی‌ثانیه هستند)

تعداد نرون	زمان ساخت پروفایل			زمان طبقه‌بندی تراکنش ورودی		
	متوسط	کمترین	بیشترین	متوسط	کمترین	بیشترین
۱۰	۷۳۶	۳۹۸	۳۰۵۹	۸۰	۵۰	۱۳۰
۲۰	۷۳۷	۴۲۰	۳۰۵۹	۸۳	۵۳	۱۳۰
۳۰	۹۰۷	۴۷۲	۳۰۴۷	۹۱	۵۹	۱۲۹
۴۰	۹۴۰	۵۰۹	۳۶۸۶	۱۰۳	۶۴	۱۵۷
۵۰	۱۰۹۶	۶۰۴	۳۷۴۷	۱۲۰	۷۶	۱۵۹

برای دستیابی به مقیاس‌پذیری، سرعت و اجرای موازی، از الگوریتم‌های نگاشت کاهشی جدول ۲ برای ساخت پروفایل و جدول ۳ برای طبقه‌بندی استفاده شد. با مقدار M برابر ۱۵ (بیشتر تعداد فرایندهای موازی) جدول ۱۱ نتیجهٔ آزمایش را نشان می‌دهد. همان‌طور که مشاهده می‌شود، هر دو زمان ساخت و طبقه‌بندی به‌طور شایان توجهی کاهش یافته‌اند. از سوی دیگر، افزایش تعداد نرون‌های شبکه، افزایش محسوسی را در زمان اجرا نشان نمی‌دهد.

جدول ۱۱. نتیجهٔ آزمون الگوریتم نگاشت کاهشی موازی (اعداد براساس میلی‌ثانیه هستند)

تعداد نرون	زمان ساخت پروفایل			زمان طبقه‌بندی تراکنش ورودی		
	متوسط	کمترین	بیشترین	متوسط	کمترین	بیشترین
۱۰	۳۷	۱۴	۴۱	۶	۵	۹
۲۰	۳۸	۱۵	۴۳	۶	۵	۱۰
۳۰	۴۱	۱۵	۴۶	۸	۵	۱۳
۴۰	۴۳	۱۶	۴۷	۹	۶	۱۵
۵۰	۴۴	۱۷	۴۹	۱۱	۶	۱۸

نتیجه‌گیری و پیشنهادها

پژوهش حاضر به پیشنهاد مدل مناسبی برای شناسایی تقلب تراکنش کارت‌های بانکی با رویکرد کشف ناهنجاری بدون سرپرستی با استفاده از شبکهٔ عصبی کوهونن و نگاشت کاهش پرداخته است. با مرور پژوهش‌های صورت‌گرفته، مشخص شد که مدل پیشنهادی با در نظر گرفتن کاستی‌هایی که در مدل‌های با سرپرستی وجود دارد، نوعی روش بدون سرپرستی ارائه داده است. همچنین با توجه به الزام سرعت زیاد پردازش در محیط بزرگ‌داده، یک رویکرد نگاشت کاهشی موازی است. بنابراین، این دو خصیصه در کنار استفاده از شبکهٔ عصبی کوهونن، نوعی نوآوری است که در ادبیات موضوع مشاهده نشد. با اجرای مدل پیشنهادی در مجموعه دادهٔ واقعی مشخص شد که رویکرد تطبیقی این مدل، در طبقه‌بندی عملکرد مناسبی را نسبت به رویکردهای کوهونن استفاده‌شدهٔ قبلی، از خود نشان داده است. درضمن، استفاده از توان تعریف وظایف موازی و متوالی نگاشت کاهش، تأثیر شایان توجهی را هم در سرعت آموزش و هم در سرعت طبقه‌بندی دارد و در مقابل بزرگ‌شدن ابعاد (تعداد نرون‌ها)، مقیاس‌پذیر است.

به‌طور خلاصه، رویکردهای تطبیقی نسبت به استفاده از مقدار برش ثابت، دقت طبقه‌بندی بیشتری دارند و در محیط بزرگ‌داده، مدل باید توان اجرای موازی مقیاس‌پذیر را داشته باشد.

نامتقارن‌بودن فضای جست‌وجو، آمارهٔ F_1 را تحت تأثیر قرار داده است و تأکیدی است بر دشواری کشف ناهنجاری در داده‌های نامتقارن که در ادبیات نیز به آن اشاره شده بود. از سوی دیگر، کنکاش در تراکنش‌هایی که عدم اعلان نادرست داشتند، نشان داد که استفاده از رویکرد خوشه‌بندی براساس رفتار تاریخی دارندهٔ کارت، نمی‌تواند مشاهداتی که مقدار ناهنجار دارند را در مجموعهٔ آموزش تشخیص دهد.

References

Axelsson, S. (2000). *The Base-Rate Fallacy and the Difficulty of Intrusion Detection*. ACM Trans .Information and System Security. 3(3): 186-205.

Bai, M., Wang, X., Xin, J. & Wang, G. (2016). *An efficient algorithm for distributed density-based outlier detection on big data*. Neurocomputing. (181): 19-28.

Bermúdez, L., Pérez, J. M., Ayuso, M., Gómez, E. & Vázquez, F.J. (2008). *A Bayesian Dichotomous, Model with asymmetric link for fraud in insurance. Insurance: Mathematics and Economics*. 42(2): 779-786.

Bhusari, V. & Patil, S. (2011). *Study of Hidden Markov Model in Credit Card Fraudulent Detection*. International Journal of Computer Applications. 20(5): 33-36.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Information Science and Statistics: Springer. Singapore.

Correa Bahnsen, A,. Stojanovic, A., Aouada, D. & Ottersten, B. (2016). Feature engineering strategies for credit card fraud detection. *Expert Systems With Applications*, 51: 134-142.

Correa Bahnsen, A,. Stojanovic, A., Aouada, D., Ottersten, B.(2014). *Improving credit card fraud detection with calibrated probabilities.* In Proceedings of the fourteenth siam international conference on data mining. Detroit, USA.

Dal Pozzolo, A., Caelen, O., Le Borgne, Y. A., Waterschoot, S. & Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications*, 41(10): 4915-4928.

Dean, J. & Ghemawat, S. (2004). MapReduce: simplified data processing on large clusters. *OSDI'04 Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation*, 137-150.

Duman, E. & Ozcelik, M.H. (2011). Detecting credit card fraud by genetic algorithm and scatter search. *Expert Systems with Applications*, 38(10): 13057-13063.

Epaillard, E. & Bouguila, N. (2016). Proportional data modeling with hidden Markov models based on generalized Dirichlet and Beta-Liouville mixtures applied to anomaly detection in public areas. *Pattern Recognition*, 55: 125-136.

González, P. C. & Velásquez, J. D. (2013). Characterization and detection of tax payers with false invoices using data mining techniques. *Expert systemps with application*s, 40(5): 1427-1436.

Halvaiee, N.S. & Akbari, M.K. (2014). A novel model for credit card fraud detection using Artificial Immune Systems. *Applied soft computing*, 24: 40-49.

Hollmén, J., Tresp, V. & Simula, O. (1999). A self-organizing map for clustering probabilistic models. *Proceedings of the Ninth International Conference on Artificial Neural Networks*, 7-10 Sept.

Huang, J., Zhu, Q., Yang, L. & Feng, J. (2016). A non-parameter outlier detection algorithm based on Natural Neighbor. *Knowledge-Based Systems*, 92: 71-77.

Huang, S. Y., Tsaih, R. H. & Yu, F. (2014). Topological pattern discovery and feature extraction for fraudulent financial reporting. *Expert systems with applications*, 41(9): 4360-4372.

Jha, S., Guillen, M. & Westland, J. C. (2012). Employing transaction aggregation strategy to detect credit card fraud. *Expert Systems with Applications*, 39(16): 12650-12657.

Kohonen, T. (1990). The self-organizing map. *In Proceedings of the IEEE*, 78 (9): 1464-1480.

Kotu, V., Deshpande, B. (2015). *Predictive analytics and data mining; Concepts and practice with RapidMiner.* Morgan Kaufmann. San Francisco.

Leskovec, J., Rajaraman, A. & Ullm, J.D. (2014). *Mining of Massive Datasets.* Cambridge University Press.

Loshin, D. (2013). *Big data analytucs; from strategic planning to enterprise integration with tools, technigues, NoSQL, and graph.* Morgan Kaufmann, London.

Minegishi, T. & Niimi, A. (2011). Proposal of Credit Card Fraudulent Use Detection by Online-type Decision Tree Construction and Verification of Generality. *International Journal for Information Security Research*, 1(4): 229-235.

Mishra, J. S., Panda, S. & Mishra, A. K. (2013). A Novel Approach for Credit Card Fraud Detection Targeting the Indian Market. *International Journal of Computer Science Issues*, 10(3): 172-179.

Nasiri, N. & Minayi, B. (2011). Data mining methods for credit card fraud detection. *1st International conference on E-Citizen & Cellphone*. Tehran: Frb 28-29. (*in Persian*)

Ngai, E. W., 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): 569-559.

Olszewski, D. (2014). Fraud detection using self-organizing map visualizing the user profiles. *Knowledge-Based Systems*, 70: 324-334.

Patidar, R. & Sharma, L. (2011). Credit Card Fraud Detection Using Neural Network. *International Journal of Soft Computing and Engineering*, (1): 32-37.

Phua, C., Lee, V., Smith, K. & Gayler, R. (2005). A Comprehensive Survey of Data Mining-based Fraud Detection Research. *Artificial Intelligence Review,* 1-14.

Powers, D. M.W. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness, correlation. *Journal of Machine Learning Technologies*, 2 (1): 37-63.

Qibei, L. & Chunhua, J. (2011). Research on Credit Card Fraud Detection Model Based on Class Weighted Support Vector Machine. *Journal of Convergence Information Technology*, 6(1): 62-68.

Quah, J. T.S. & Sriganesh, M. (2008). Real-time credit card fraud detection using computational intelligence. *Expert Systems with Applications*, 35(4): 1721-1732.

Raj, B. E. & Portia, A. (2011). Analysis on Credit Card Fraud Detection Methods. *International Conference on Computer,* Communication and Electrical Technology, March.

Rama Kalyani, K. & Uma Devi, D. (2012). Fraud Detection of Credit Card Payment System by Genetic Algorithm. *International Journal of Scientific & Engineering Research,* 3(7): 1-6.

Rojas, R. (1996). *Neural Networks.* Berlin: Springer-Verlag.

Srivastava, A., Kundu, A., Sural, S. & Majumdar, A. K. (2008). Credit Card Fraud Detection Using Hidden Markov Model. *IEEE transactions on dependable and secure computing*, 5(1): 37-48.

Swets, J.A. (1996). *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers*. Lawrence Erlbaum Associates, Mahwah, N.J.

Tripathi, K.K. & Ragha, L. (2013). Hybrid Approach for Credit Card Fraud Detection. *International Journal of Soft Computing and Engineering*, 3(4): 8-11.

Vosough, M., Taghavifard, M.T. & Alborzi, M. (2015). Bank card fraud detection using artificial neural network. *Journal of Information Technology Management,* 6(4): 721-746. (*in Persian*)

Whitrow, C., Hand, D.J., Juszczak, P. & Weston, D. (2008). Transaction aggregation as a strategy for credit card fraud detection. *Data mining and knowledge discovery*, 18(1): 30-55.

Zareapoor, M., Seeja, K.R. & Alam, M. A.(2012). Analysis of credit card fraud detection techniques: based in certain design criteria. *International journal of computer applications,* 52(3): 35-42.

Zareapoor, M. & Shamsolmoali, P. (2015). Application of credit card fraud detection: based on bagging ensemble classifier. *Procedia computer science*, 48: 679-685.

Zaslavsky, V. & Strizhak, A. (2006). Credit card fraud detection using self organizing maps. *International Journal of Information & Security*, 18: 48-63.