ارائه مدل شناسایی متقلبین در سیستم‌های بانکداری آنلاین بر مبنای تراکنش‌های کارت‌های اعتباری با استفاده از جنگل تصادفی وزن‌دار چندگانه و مدل کوادراتیک

چکیده

با رشد روزافزون بانکداری برخط[1] بانک‌ها و مؤسسات مالی روزبه‌روز بیشتر به سمت استفاده از این فناوری و خدمات آن سوق پیدا می‌کنند. باتوجه‌به حجم بالای تراکنش‌ها امکان مدیریت آن‌ها توسط نیروی انسانی عملاً غیرممکن است. به همین منظور امروزه رویکردهای مبتنی بر داده‌کاوی به کمک بانکداری برخط آمده است. در این مقاله یک مدل کارآمد برای شناسایی متقلبین در تراکنش‌های کارت‌های بانکی ارائه می‌گردد. روش پیشنهادی از ماتریس مجاورت، جای‌گذاری ویژگی‌های بدون مقدار با استفاده از وزن‌دهی و الگوریتم تجمیعی جنگل تصادفی استفاده می‌کند که در هر انشعاب آن با محاسبه وزن هر انشعاب، بهترین انشعاب تصمیم‌گیرنده با استفاده از محاسبه هزینه مدل انتخاب می‌شود. همچنین با استفاده از مدل کوادراتیک چندگانه بهترین جنگل را برای تصمیم‌گیری انتخاب می‌نماید. بدین ترتیب این روش را بر روی دو مجموعه‌داده که اولی ۱۴ ویژگی و دومی ۲۰ ویژگی داشته است تست کرده‌ایم و مشاهده شده است که مدل این تحقیق در مقایسه با درخت تصمیم و ماشین بردار پشتیبان و شبکه عصبی و جنگل تصادفی معمولی که در حال حاضر بالاترین نتایج را نسبت به هر روشی از خود نشان داده‌اند نیز بهبودهایی داشته است. همچنین آزمایشات نشان می‌دهد که هیچ یک از روش‌های مذکور قادر به پیش‌بینی خطای OOB نبوده و جنگل تصادفی معمولی که قادر به پیش‌بینی این خطا می‌باشد بسیار ضعیف‌تر از مدل پیشنهادی عمل نموده است. فقط روش پیشنهادی است که می‌تواند این مقدار را محاسبه نماید و مقدار مناسبی برای آن پیش‌بینی نماید. همچنین در آزمایش روی مجموعه‌داده‌ی دوم نیز در همین حدود بهبودهایی داشته ایم که به تفضیل در مقاله ذکر شده است.

کلمات کلیدی: شناسایی متقلبین، بانکداری آنلاین، تراکنش‌های کارت‌های اعتباری، ماتریس مجاورت، الگوریتم تجمیعی، جنگل تصادفی وزن‌دار چندگانه، مدل کوادراتیک

[1] online

مقدمه

مزایای بانک‌داری آنلاین را می‌توان از دو جنبه مشتریان و مؤسسات مالی موردتوجه قرار داد. از دید مشتریان می‌توان به صرفه‌جویی در هزینه‌ها، صرفه‌جویی در زمان و دسترسی به کانال‌های متعدد برای انجام عملیات بانکی نام برد. از دید مؤسسات مالی می‌توان به ویژگی‌هایی چون ایجاد و افزایش شهرت بانک‌ها در ارائه نوآوری، حفظ مشتریان علی‌رغم تغییرات مکانی بانک‌ها، ایجاد فرصت برای جستجوی مشتریان جدید در بازارهای هدف، گسترش محدوده جغرافیایی و فعالیت و برقراری شرایط رقابت کامل را نام برد (ویشال و همکاران ۲۰۲۱).

امروزه بانک‌ها و مؤسسات مالی و اعتباری برای خدمات‌رسانی مؤثر، ناگزیر از مهاجرت از بانک‌داری سنتی به بانک‌داری مدرن و برخط شده‌اند. هر چند استفاده از این سامانه‌ها باعث مدیریت بهتر فرایندهای مالی و افزایش کارایی و سرعت خدمات‌رسانی به مشتریان این مؤسسات شده، اما تقلب و سوءاستفاده‌های مالی یکی از مشکلاتی است که این سازمان‌ها در پی پیشگیری از آن‌ها و کاهش اثرات آن‌ها بوده‌اند. ایجاد سیستم‌های کارا برای شناسایی مشتریان بانک و کنترل فعالیت‌های مالی آنها بهترین روش برای مبارزه با مجرمان در جریان انجام عملیات بانکی است (فانگ و همکاران ۲۰۲۱).

امروزه روش‌های داده‌کاوی به‌عنوان بهترین راهکار برای شناسایی خودکار تقلب در حوزه‌های مختلف شناخته شده‌اند. داده‌کاوی به‌عنوان فرایند کشف و استخراج الگوهای پنهان از حجم بالایی از داده‌ها تعریف می‌شود. در سامانه‌های بسیاری از روش‌های داده‌کاوی برای شناسایی و کشف تقلب و سوءاستفاده مالی استفاده شده است (پتیدار و شارما ۲۰۱۱).

بخش عمده‌ای از فعالیت‌های متقلبانه، معطوف به تراکنش با کارت‌های اعتباری است. از این رو ایجاد سیستمی که ناظر بر عملکرد نظام‌های پرداخت باشد، به‌منظور شناسایی تقلب در تراکنش‌های موجود در کارت‌های اعتباری بانکی، ضروری به نظر می‌رسد. تاکنون مطالعات مختلفی برای شناسایی تقلب در تراکنش‌ها انجام شده است که هر کدام دارای مزایا و معایب خاص خود هستند. در این مقاله روشی ارائه می‌شود که علاوه بر تعیین ویژگی‌های مؤثر در دسته‌بندی تراکنش‌ها، از دقت و سرعت عملکرد بهتری نسبت به سایر روش‌ها، بهره‌مند گشته است(کارتا و همکاران در ۲۰۲۰). ازآنجایی که روش های داده کاوی به‌صورت پویا با محیط های در حال تغییر، سازگار می شوند، مدت زمانی را که صرف کشف الگو می شود، تا حد قابل توجهی نسبت به روش های غیر خودکار کاهش می دهند(وانگ و همکاران ۲۰۲۱).

در این تحقیق به دنبال پاسخ به این سؤال هستیم که استفاده از ماتریس مجاورت و الگوریتم تجمیعی جنگل تصادفی، کارایی کشف تقلب در سیستم بانکی را تا چه حد بهبود می‌بخشد؟

مبانی نظری و پیشینه تحقیق

در مقاله قلی‌پور و همکاران به‌منظور شناسایی تقلب در تراکنش‌های موجود در کارت‌های اعتباری بانکی، از روش یادگیری دسته‌جمعی استفاده گردیده است. در این روش علاوه بر تعیین ویژگی‌های مؤثر در دسته‌بندی تراکنش‌ها، دقت و صحت دسته‌بندی افزایش یافته است (قلی‌پور و همکاران ۱۴۰۰).

در مقاله وثوق و همکاران به روش‌های کشف تقلب پرداخته شده است. در تقلب‌های برخط، تراکنش‌ها از راه دور انجام شده و تنها به جزئیات کارت نیاز است نه لزوماً خود کارت. آن‌ها دریافتند که به علت استفاده‌ی گسترده از اینترنت، کاربران می‌توانند موقعیت و هویت تراکنش اینترنتی خود را پنهان کنند (وثوق و همکاران ۱۳۹۸).

در مقاله حاتمی راد و همکاران به موضوع تقلب و اینکه در تقلب قصد شخصی و فریب دیگران مطرح است می‌پردازد. روش مطرح شده در این مقاله به استخراج ویژگی‌های پنهان پرداخته و سپس از بین ویژگی‌های استخراجی بهترین ویژگی‌ها جهت تصمیم‌گیری در مورد یک تراکنش را انتخاب می‌نماید (حاتمی راد و همکاران ۱۳۹۷).

بنایی و همکاران در مقاله خود مطرح نمودند که باتوجه‌به حجم بالای تراکنش‌ها باید با رویکردهای مبتنی بر داده‌کاوی به کمک بانکداری برخط رفت تا مشکلات مربوط به تقلب را شناسایی نمود. آن‌ها از روش پردازش داده‌های حجیم بهره بردند (بنایی و همکاران ۱۳۹۶).

نظیر و همکارانش در ۲۰۲۳ مقاله‌ای در راستای تشخیص تقلب در کارت‌های اعتباری با استفاده از دو تکنیک یادگیری عمیق و رمزنگاری اطلاعات منتشر کردند. آن‌ها در لایه ابتدایی داده‌ها را رمزگذاری و در لایه انتهایی داده‌ها را رمزگشایی نمودند. اگرچه این روش پیچیدگی محاسباتی را افزایش داده است اما استفاده از رمزنگاری در کنار یادگیری عمیق چندلایه سبب افزایش دقت و قدرت تشخیص تقلب و همچنین کاهش احتمال وقوع تقلب شده است..

موریرا و همکاران در ۲۰۲۲ روشی مبتنی بر رگرسیون لجستیک، بیزین و نزدیک‌ترین همسایه برای شناسایی تقلب در سیستم بانکی مطرح نمودند. آن‌ها ابتدا به استخراج ویژگی‌های مؤثر پرداختند و فضای ویژگی ایجاد شده را با روش‌های نام برده شده دسته‌بندی نمودند. از نظر نویسندگان این مقاله، روش‌های یادگیری ماشین قادر

است با دقت بالایی تراکنش‌های بانکی را دسته‌بندی نماید و علاوه‌برآن می‌تواند صحت دسته‌بندی تراکنش‌ها را نیز افزایش دهد. نتایج به دست آمده نیز بیانگر همین موضوع است.

در مقاله دیگر مجموعه‌داده آموزش جدید، برای تبدیل داده بانک به داده‌های مناسب برای الگوریتم CLOPE که در خصوص تکنیک خوشه‌بندی برای داده اسمی (مقادیر رشته‌ای) می‌باشد، به‌منظور تشخیص موارد تقلب، ارائه گردیده است. نتایج آزمایشی نشان می‌دهد که CLOPE یک الگوریتم مناسب برای تشخیص موارد متقلبانه می‌باشد. اما این سیستم نمی‌تواند به‌تنهایی، به طور کامل، اجرا گردد و باید از توانائی تحلیل‌گران در تجزیه‌وتحلیل داده‌ها، و ارائه مجموعه‌ای از قوانین (معیارهای تعیین شده) برای اعتباربخشی به خوشه‌ها پس از عمل خوشه‌بندی، استفاده نمود (ویشال و همکاران ۲۰۲۱).

مقاله (دزاسکس و همکاران ۲۰۲۱) یک رویکرد تشخیص ناهنجاری هیبریدی است که استفاده از خوشه‌بندی برای ایجاد رفتارهای نرمال مشتریان و استفاده از تکنیک‌های آماری برای تعیین انحراف معامله خاص از رفتار گروه مربوطه است. این رویکرد بر روی یک مجموعه‌داده واقعی که شامل ۸.۲ میلیون معاملات انجام می‌شود، مورد آزمایش قرار گرفته و نتایج نشان می‌دهد که TEART به‌خوبی از لحاظ پارتیشن‌هایی که در مقایسه با الگوریتم K-mean سنتی به دست می‌آید، خوب است.

در (آتا و همکاران ۲۰۲۱) نویسندگان برای تشخیص رفتارهای غیرمعمول مشتری ماشین بردار پشتیبان (هند و همکاران ۲۰۱۹) را توسعه داده‌اند. آن‌ها ترکیبی از الگوریتم‌های نظارت شده و بدون ناظر ماشین بردار پشتیبان را ارائه داده‌اند. مزیت این روش آن است که می‌تواند با مجموعه‌داده‌های ناهمگون کار کند. بااین‌حال ارزیابی عملکرد آن بر اساس مجموعه‌داده‌های شبیه‌سازی‌شده برای موارد مشکوک می‌باشد.

نویسندگان در (فوآ ۲۰۲۰) بیان داشته‌اند، درخت تصمیم یکی از پرکاربردترین روش‌های استنباط استقرایی از سال ۱۹۶۰ تاکنون است. در این مقاله روش درخت تصمیم برای تعیین میزان ریسک تشخیص تقلب و پول‌شویی، بر اساس مشخصات مشتری بکار گرفته شده است

در طرح (یونکوک و همکاران ۲۰۱۹)، یک چارچوب تشخیص تقلب آنلاین بانکی مؤثر ارائه شده که به ترکیب منابع مربوط و به‌کارگیری تکنیک‌های داده‌کاوی پیشرفته می‌پردازد. در با ایجاد یک بردار تقابل[2] برای هر تراکنش، بر اساس توالی رفتار مشتری در طول زمان، نرخ تمایز تراکنش جاری مشتری را با رفتار رایج وی

[2] Contrast Vector

نشان داده شده است. در این تحقیق، یک الگوریتم جدید برای استخراج مؤثر الگوهای تضاد و تشخیص رفتار جعلی از رفتار واقعی، برگرفته از یک الگوی انتخاب مؤثر و رتبه‌بندی ریسک که پیش‌بینی‌ها از مدل‌های مختلف را ترکیب می‌کند، معرفی شده است.

در سال‌های اخیر (سالچنبرگر و همکاران ۲۰۱۸)، شبکه عصبی مصنوعی (ANN) را به دلیل زمینه‌های کاربرد گسترده مطرح نمودند. در بسیاری از این برنامه‌های کاربردی تمرکز بر یادگیری حساس به هزینه وجود دارد به‌طوری‌که هزینه‌های مختلفی برای انواع مختلف طبقه‌بندی نادرست وجود دارد. هزینه طبقه‌بندی نادرست یک مثال از یک زمینه متفاوت است. در بسیاری از طبقه‌بندی‌های دودویی حساس به هزینه مانند مشکلات تشخیص، دو طبقه‌بندی نادرست مختلف وجود دارد و هر یک از آنها دارای هزینه است. بااین‌حال، در کسب‌وکار مشکلاتی از قبیل تشخیص جعل کارت اعتباری و بازاریابی مستقیم هر مشاهده طبقه‌بندی شده، می‌تواند هزینه متفاوتی داشته باشد و علاوه بر این ممکن است سودی برای درستی طبقه‌بندی هر یک وجود داشته باشد؛ بنابراین، در چنین مواردی، ضرورتی برای توسعه یک مدل طبقه‌بندی وجود دارد که به منافع و هزینه‌های شخصی رسیدگی کند.

در مقاله‌ای دیگر (واهسلر و همکاران ۲۰۱۵) نشان داده شده است که استفاده از قواعد انجمنی و به‌کارگیری آن‌ها بر روی مجموعه‌دادگان تراکنش‌های بانکی نتایج مطلوبی را حاصل نموده است. همچنین در رویکرد دیگری (پوزولو و همکاران ۲۰۱۴ و باتاچاریان و همکاران ۲۰۱۱) ترکیب این روش در کنار تکنیک یادگیری ماشین توانسته است دقت و صحت تشخیص را افزایش دهد.

نویسندگان در (باتاچاریا و همکاران ۲۰۱۱)، به بررسی برخی مدل‌های پیش‌بینی‌کننده‌ی معروف داده‌کاوی برای شناسایی تقلب پرداخته‌اند. در این مطالعه از دو تکنیک داده‌کاوی رگرسیون لجستیک و ماشین بردار پشتیبان برای شناسایی تقلب استفاده شده است. در ادامه ابتدا به معرفی این دو تکنیک می‌پردازیم.

هنگام برخورد با ساختارهای گرافی، ناهنجاری می‌تواند طبق مشخصات گراف به‌خوبی طبقه‌بندی شود. در تحقیق (ابرله ۲۰۰۷) ناهنجاری‌ها را بر اساس آنچه گفته شد به سه گروه تقسیم‌بندی کرده‌اند. انجام درج با وجود یک رأس یا یک لبه غیرمنتظره در گراف. انجام اصلاح با حضور یک برچسب غیره منتظره روی یک رأس یا یک لبه. حذف شامل عدم وجود یک رأس یا لبه‌ی مورد انتظار می‌باشد. گاهی اوقات، حتی شامل مفاهیم، مرتبط با لبه نیز می‌شود به‌عنوان‌مثال حذف یک رأس خاصی از کل لبه‌های مجاور که ممکن است حتی حذف شده باشد. این امر سبب بهبود تشکیل گراف و کشف ناهنجاری در یال‌های آن می‌شود. این یال‌ها هریک می‌تواند یک تراکنش را شامل شود.

نویسندگان در (شن و همکاران ۲۰۰۷)، برای شناسایی تقلب در کارت‌های اعتباری، مقایسه‌ای بین روش‌های داده‌کاوی مختلف ارائه داده‌اند. در این مطالعه، سه روش پرکاربرد نظیر درخت تصمیم، شبکه‌ی عصبی و رگرسیون لجستیک استفاده شده است.

کارکردهای مدیریتی:

با استفاده از روش ارائه شده در این مقاله و زمانی که تراکنش های مشکوک شناسایی شوند مدیریت تراکنش ها بسیار ساده تر خواهد شد و هزینه های مدیریتی کاهش پیدا خواهد کرد.

شکاف تحقیقاتی

شکاف عملی

تراکنش‌های تقلبی هزینه های هنگفتی را به بانک ها تحمیل می‌کنند. بنابراین مدلی مناسب و سود ده است که هزینه‌ها را تا حد امکان کاهش دهد. داده‌های بانک ها حجم بالایی دارند و پردازش آن‌ها زمان بر است در نتیجه آن‌ها ملزم به استخدام نیروی انسانی زیادی هستند که زمان بر است وهزینه بالایی با به بانک تحمیل می‌کند.

شکاف نظری

باتوجه‌به بررسی مطالعات انجام شده، شکاف تحقیقاتی موجود، عدم وجود مدلی است که بتواند اطلاعات حجیم و غیرنرمال را کاوش نموده و روشی مؤثر برای استخراج اطلاعات ارزشمند در داده‌های بزرگ و داده-های ناهمخوان باتوجه‌به هزینه مدل باشد. مدلی که قادر باشد هم‌زمان با فرایند آموزشش، متغیرهای مهمی که بیش‌ترین تأثیر را در طبقه‌بندی داده‌ها دارند، استخراج کند و سرعت پاسخگویی را افزایش دهد و بهترین مدل را با احتمال و دقت بالا تشخیص دهد. همچنین بتواند داده‌های بدون مقدار یا دارای مقدار نامتعارف و پرت را شناسایی نموده و برای آن‌ها مقدار مناسب جایگزین نماید.

روش تحقیق

تحقیق حاضر یک تحقیق کاربردی در حوزه بانکداری آنلاین می‌باشد. مدل مطرح شده در این تحقیق می‌تواند به بهبود کیفیت خدمات بانکی و امنیت تراکنش‌ها کمک نماید.

در این مقاله برای شناسایی متقلبین، از ماتریس مجاورت، مدل تجمیعی جنگل تصادفی وزن‌دار و الگوریتم مدل احتمالاتی استفاده کرده‌ایم. در مدل پیشنهادی ویژگی‌های بدون مقدار و دارای مقدار غیرنرمال مقداردهی مناسب می‌گردد. همچنین با استفاده از جنگل تصادفی وزن‌دار که برای هر انشعاب آن وزن و هزینه محاسبه می-

شود و برای هر دسته داده بهترین انشعاب جهت تصمیم‌گیری انتخاب می‌گردد، بهترین و کم‌هزینه‌ترین تصمیم برای هر نوع داده اتخاذ می‌شود. همچنین با استفاده از الگوریتم محاسبه احتمال، بهترین مدل بر روی هر انشعاب ایجاد می‌گردد. در واقع این مدل یک روش یادگیری دسته‌جمعی محسوب شده و برای یادگیری از تعداد زیادی درخت تصمیم استفاده می‌کند که در هر درخت محاسبه وزن و هزینه برای هر نوع داده و هر ویژگی لحاظ می‌شود. پس از پایان آموزش، برای دسته‌بندی یک نمونه‌ی جدید، بین درختان رأی‌گیری بر اساس بهترین مدل برای آن داده انجام شده و کلاس با بیش‌ترین رأی و کم‌ترین هزینه، برای نمونه‌ی جدید انتخاب می‌شود.

در این مدل، دو پارامتر توسط کاربر تعیین می‌شود. پارامتر اول تعداد درختان تصمیم است که در جنگل ساخته خواهد شد. این پارامتر بسته به تعداد داده‌های آموزشی و تعداد ویژگی‌های موجود در هر مسئله‌ای متغیر است.

پارامتر دوم با نام m شناخته می‌شود. در زمان انتخاب ویژگیِ شکست در هر گره‌ی درخت، باید m ویژگی به‌صورت تصادفی انتخاب شده و از بین این m ویژگی، با کمک معیارهای کارایی بهترین ویژگی برای بخش‌بندی داده‌ها انتخاب شود.

این پارامتر در کل فرایند آموزش مدل و در بین کلیه‌ی درختان ثابت در نظر گرفته می‌شود. مقادیر معمولی که برای این پارامتر در نظر گرفته می‌شود، عبارت‌اند از: Sqrt(nVariable)، Log(nVariable)+1 و منظور از nVariable، تعداد ویژگی‌های موجود در مجموعه‌داده‌ها یا همان متغیرهای مسئله است.

همچنین برای آموزش هر درخت، از مجموعه‌داده‌های آموزشی اولیه، به طور تصادفی و با جای‌گذاری به تعداد N نمونه‌ی آموزشی انتخاب می‌شود. پارامتر N را معمولاً به اندازه‌ی کل مجموعه‌داده‌های اولیه‌ای که در دسترس است، در نظر می‌گیرند.

بنابراین، ممکن است در بین زیرمجموعه‌های آموزشی ایجاد شده اشتراکاتی وجود داشته باشد. نکته‌ی قابل‌توجه اینجاست که این مدل می‌تواند پدیده‌ی بیش‌برازش[۳] را به‌خوبی مدیریت کرده و عمومیت بیش‌تری را در فضای مسئله‌اش داشته باشد. منظور از عمومیت این است که مدل در مواجهه با داده‌های جدید، بتواند به‌خوبی با آن‌ها تطبیق پیدا کرده و کمترین خطای تشخیص یا پیش‌بینی را در پی داشته باشد. یعنی تنها به داده‌های آموزش محدود نبوده (overfit نشده) و عمومیت داشته باشد. در واقع به دلیل همین سیاست تصادفی عمل‌کردن در انتخاب زیرمجموعه‌های آموزشی و به‌خصوص در انتخاب m ویژگی، به‌خوبی از پدیده‌ی بیش‌برازش جلوگیری می‌شود.

[3] overfit

روند ساخت مدل

مدل پیشنهادی از ۳ بخش اصلی تشکیل می‌شود.

در بخش اول بارگذاری و پیش‌پردازش داده‌ها انجام می‌شود. مجموعه‌داده‌های در دسترس شامل ویژگی‌های عددی پیوسته و ویژگی‌های رده‌ای می‌باشند.

پس از بارگذاری داده‌ها، اگر ویژگی‌های رده‌ای با مقادیر رشته‌ای مقداردهی شده باشند، باید این مقادیر رشته-ای به عدد تبدیل شوند.

بخش دوم، پارامترهای موردنیاز الگوریتم مقداردهی اولیه می‌شوند. این پارامترها عبارت‌اند از: تعداد متغیرهای مسئله یا همان ویژگی‌های موجود در داده‌ها، تعداد نمونه‌های موجود در مجموعه‌داده‌ها، تعداد درختان موردنیاز، تعیین پارامتر N که تعداد نمونه‌های آموزشی را برای هر درخت تعیین می‌کند و تعیین پارامتر m برای مشخص‌شدن تعداد ویژگی‌هایی که در هر گره باید بررسی شده و بهترین آن برای بخش‌بندی داده‌ها انتخاب شود.

بخش سوم، مدل ساخته می‌شود. برای ساخت مدل، به‌ازای هر درخت، به تعداد N نمونه به‌صورت تصادفی و با جای‌گذاری از مجموعه‌داده‌ی اولیه انتخاب و به‌عنوان trainingSample ذخیره می‌شود.

روند کلی ساخت مدل در ادامه شرح داده شده است.

۱. به تعداد درختان تصمیم موردنظر، از داده‌های آموزشی اولیه، زیرمجموعه داده‌های آموزشی را نمونه-گیری می‌کنیم.
۲. در هر زیرمجموعه‌ی آموزشی، یک درخت تصمیم را بدون هرس‌کردن و تا انتهای فرایند آموزشش توسعه می‌دهیم. برای هر درخت باتوجه‌به تعداد ویژگی‌ها، یک وزن اولیه در نظر می‌گیریم. در هر گره، به‌جای انتخاب بهترین ویژگیِ شکست از بین کلیه‌ی ویژگی‌ها، m ویژگی را به طور تصادفی انتخاب کرده و از بین این m ویژگی، بهترین را به‌عنوان ویژگی شکست انتخاب می‌کنیم. وزن انشعاب را باتوجه‌به اهمیت ویژگی انتخاب شده و هزینه محاسبه آن برای نمونه‌های موجود در انشعاب به‌روز می-نماییم. یادگیرنده‌های ضعیف در حین اضافه‌شدن به مجموعه، وزن‌دهی می‌شوند که این وزن‌دهی بر اساس میزان دقت در طبقه‌بندی نمونه‌هاست. پس از اضافه‌شدن هر طبقه‌بند، نمونه‌های موجود (داده‌ها) نیز وزن‌دهی می‌گردند (وزنشان اصلاح می‌گردد). وزن‌دهی نمونه‌ها به صورتی است که در هر مرحله،

وزن نمونه‌هایی که به‌صورت صحیح طبقه‌بندی می‌شوند کاهش یافته و وزن نمونه‌هایی که به‌درستی طبقه‌بندی نشده‌اند، بیشتر می‌شود تا در مراحل بعدی (توسط یادگیرنده‌های جدید) بیشتر موردتوجه بوده و با دقت بیشتری طبقه‌بندی گردند؛ بنابراین تمرکز یادگیرنده‌های ضعیف جدید، بیشتر بر روی داده‌های خواهد بود که سیستم در مراحل قبلی قادر به طبقه‌بندی صحیح آنها نبوده است.

۳. ماتریس هزینه داده‌ها را محاسبه می‌نماییم.

در جدول ۱ ماتریس هزینه محاسبه شده نشان داده شده است.

جدول ۱: ماتریس هزینه برای دسته بندی

پیش ینی نادرست	c(i,i) یا TN	c(i,i) یا FN	1	2
پیش‌بینی درست	c(i,i) یا FP	c(i,i) یا TP	1	0

لازم به ذکر است که c(i,i) که همان (TN و TP) می‌باشد هنگامی که نمونه به‌درستی پیش‌بینی شود معمولاً به‌عنوان سود در نظر گرفته می‌شود. همچنین اقلیت یا طبقه نادر به‌عنوان کلاس مثبت در نظر گرفته می‌شود. مشخص است که هزینه‌ی دسته‌بندی نادرست یک نمونه‌ی اقلیت بیشتر از هزینه‌ی دسته‌بندی نادرست یک نمونه‌ی اکثریت است به همین دلیل ارزش FN معمولاً بیشتر از FP است.

۴. هزینه مورد انتظار نمونه‌ها محاسبه می‌شود.

باتوجه‌به ماتریس هزینه یک نمونه باید در رده‌ای که کمترین هزینه از آن انتظار می‌رود دسته‌بندی شود. هزینه مورد انتظار R(i|x) از دسته‌بندی نمونه x در رده i ام به شرح ذیل بیان می‌گردد.

فرمول ۱:

$$R(i|x) = \sum_j P(j|x)C(i,j) \quad .۱$$

که در آن $P(j|x)$، احتمال تعلق نمونه x به رده ی j را مشخص می‌کند. در ابتدا هر دسته برای هر مجموعه به طور مستقل استفاده می‌شود. سپس نتایج به‌منظور تولید با هم ترکیب می‌شوند. روش مبتنی بر ترکیب دسته-بندی‌ها معمولاً دارادی دقت بالاتر، همچنین FP کمتر نسبت به دسته‌بندی‌های جداگانه دارد.
روند کلی روش پیشنهادی در ادامه شرح داده شده است.

۵. هر انشعاب را توسعه می‌دهیم تا زمانی که آموزش آن به پایان رسد.

داده‌ی جدید را به کلیه‌ی درختان اعمال کرده و برچسب نهایی داده، رأی اکثریت درختان و درنظرگرفتن کمترین هزینه خواهد بود. در مسئله‌ی رگرسیون، مقدار پیش‌بینی‌شده‌ی نهایی، میانگین مقادیر پیش‌بینی شده توسط همه‌ی درختان خواهد بود. روش‌های مختلفی برای ترکیب نتایج دسته‌بندی‌کننده‌ها وجود دارد، متداول‌ترین روش‌ها میانگین‌گیری و یا استفاده از رأی اکثریت هستند. به‌منظور نشان‌دادن رویکرد ترکیبی از دو تکنیک استفاده می‌شود. اولین روش قانون رأی‌گیری اکثریت[4] می‌باشد. در این روش اظهارنظر هر دسته‌بند در مورد کلاس الگوی ورودی، به‌عنوان یک رأی محسوب می‌شود و تصمیم‌گیری نهایی بر اساس آرای اخذ شده از دسته‌بندی‌های مختلف صورت می‌گیرد. در این تکنیک تمامی داده‌ها دارای وزن یکسان هستند و حساسیتی نسبت به هزینه داده‌ها وجود ندارد.

۶. بهترین مدل برای داده محاسبه می‌گردد.
فرض ابتدایی در تعیین بهترین مدل این است که داده‌ها به‌صورت نرمال توزیع شده‌اند. در این روش نیازی به برابر بودن کوواریانس بین دو گروه و انشعاب نیست. ابتدا مجموعه‌ای از مشاهدات به نام x به طبقه‌بند داده می‌شوند که از نوع خانواده y هستند. به این قسمت مرحله تعلیم گفته می‌شود که متناسب با آن طبقه-بند، سطح فرضی را برای جداسازی دو دسته از ویژگی‌ها بکار می‌برد. سپس پس از تعلیم طبقه‌بند، داده‌ها تست به آن داده می‌شود تا دو دسته ویژگی‌های مختلف را از یکدیگر جدا کند.

در روش پیشنهادی برای حساسیت به هزینه از تابع باور و برای محاسبه احتمال از تابع کوادراتیک استفاده می‌شود. به‌این‌ترتیب که زمانی می‌توانیم یک تراکنش را قانونی بدانیم که هر دو دسته‌بندی‌کننده آن را

4. Majority voting rule.

قانونی تشخیص دهند. با این کار می‌توان معیار FN (تراکنش ورودی تقلبی می‌باشد و سیستم آن را به‌اشتباه قانونی تشخیص داده است) که بیشترین هزینه را برای ما دارد به حداقل رساند. همچنین برای ایجاد وزن برای نمونه‌ها از تکنیک میانگین قانون[5] استفاده می‌نماییم. به این صورت هر نمونه را باتوجه‌به الگوی ورودی، با یک میانگین احتمال خلفی به یک کلاس تخصیص می‌دهیم. با استفاده از الگوریتم کوادراتیک، ویژگی‌های حاصل شده برای هر مدل و انشعاب طبقه‌بندی می‌شوند.

ماتریس مجاورت[6]

یک ماتریس N*N است و N تعداد کل مجموعه‌داده‌ی آموزشی اولیه است (بنسال و شرما ۲۰۲۱). زمانی که یک درخت تصمیم ساخته شد، داده‌های آموزش متعلق به خودش را به درخت اعمال می‌کنیم. اگر نمونه‌ی i ام با نمونه‌ی j ام در یک نود پایانی مشابه قرار گرفتند، عنصر (i,j) ماتریس مجاورت را یکی اضافه می‌کنیم. در نهایت، درایه‌های ماتریس را با تقسیم‌کردن بر تعداد کل درختان، نرمال می‌کنیم.

ماتریس مجاورت می‌تواند در تعریف ساختار داده‌ها و یا یادگیری غیر نظارت شده به کار گرفته شود (بنسال و شرما ۲۰۲۱).

ویژگی‌های مدل پیشنهادی

در نهایت، ویژگی‌های مدل پیشنهادی را به‌صورت زیر خلاصه می‌کنیم:

- یک الگوریتم قدرتمند با ماهیت دسته‌جمعی بودن، محسوب شده که قدرت یادگیری و تعمیم خوبی را فراهم می‌آورد.
- این الگوریتم به‌صورت کارا بر روی مجموعه‌داده‌های بسیار بزرگ اجرا می‌شود.
- بدون حذف ویژگی (یا متغیر) می‌تواند با بیش از هزار متغیر در مسئله کار کند.
- برآوردی از میزان اهمیت هر یک از متغیرها را نشان می‌دهد.
- در برابر پدیده‌ی بیش‌برازش مقاوم است.
- می‌تواند یک تخمین بدون بایاس داخلی از خطای عمومی‌سازی را در حین فرایند آموزش، ارائه دهد.

5.Average rule
[6]Proximity Matrix

- برای افزایش سرعت و کاهش زمان فرایند آموزش، می‌توان به‌صورت موازی درختان را آموزش داد.
- دارای روشی مؤثر برای برآورد داده‌های گم شده است و بدون کاهش دقت می‌تواند در مجموعه‌داده‌هایی که مقادیر گم‌شده‌ی زیادی دارند، به‌خوبی کار کند.
- مدل‌های ساخته شده می‌توانند برای استفاده‌های بعدی بر روی داده‌های دیگر، ذخیره شوند.
- این الگوریتم میزان مجاورت (یا شباهت) بین هر جفت از داده‌ها را محاسبه می‌کند. این امکان می‌تواند در خوشه‌بندی، برآورد داده‌های گم شده و همچنین ایجاد دید کلی در مورد داده‌ها بسیار مؤثر باشد.

نتایج روش پیشنهادی

برای بررسی تأثیر پارامترهای مختلف مدل پیشنهادی آزمایشاتی را طراحی و اجرا کرده‌ایم. با این آزمایش‌ها مقادیر قابل‌قبولی برای پارامترهای آزاد مدل تعیین می‌شود. در ادامه نیز برای نمایش قدرت و کارایی مدل پیشنهادی، آن را با سایر روش‌های موجود مقایسه و نتایج به دست آمده ارائه شده است.

داده‌ها و نرم‌افزار مورداستفاده

در این تحقیق از دو مجموعه‌داده‌ی استاندارد، متعلق به کارت‌های اعتباری موجود در کشور آلمان و کشور استرالیا استفاده می‌نماییم. در هر دو مجموعه‌داده برای حفظ امنیت، نام و مقادیر فیلدها به‌صورت یکتا، گمنام[۷]-سازی شده‌اند. داده‌های فیلتر شده شامل نام و نام خانوادگی مشتریان و اطلاعات شخصی آن‌ها می‌باشد.

پارامترهای منحصربه‌فرد در مدل پیشنهادی

به‌طورکلی در هر الگوریتم تعدادی پارامتر آزاد و مؤثر در کارایی الگوریتم وجود دارد. الگوریتم پیشنهادی نیز از این قضیه مستثنی نبوده و تعدادی پارامتر دارد که تنظیم بهینه‌ی آن‌ها موجب افزایش دقت و قدرت الگوریتم در شناسایی تقلب خواهد شد. در این بخش در قالب آزمایشاتی تأثیر این پارامترها را نشان داده‌ایم.

کلیه‌ی آزمایش‌ها انجام شده در این بخش بر روی دو مجموعه‌داده‌ی استانداردی که در بخش قبل معرفی شد، انجام شده و نتایج ارائه شده‌اند.

[7] Anonymous

تعداد انشعابات

یکی از پارامترهای مهم و مؤثر در دقت شناسایی، تعداد انشعابات در مدل پیشنهادی است. هر یک از انشعابات موجود در مدل به‌تنهایی کارایی زیادی نداشته و در واقع قدرت تشخیص مدل به برآیند قدرت کلیه‌ی آن‌ها وابسته است. در واقع این ویژگیِ روش‌های دسته‌جمعی است که از نوعی سیستم رأی‌گیری بین اعضا برای تصمیم‌گیری نهایی استفاده می‌کند. هر چه تعداد این انشعاب‌ها در مدل بیش‌تر باشد دقت تشخیص بهبود می‌یابد.

برای مشاهده‌ی تأثیر این پارامتر در کارایی الگوریتم و انتخاب تعداد مناسب آن در مدل، آزمایش زیر انجام شده است:

آزمایش ۱: تأثیر پارامتر تعداد انشعابات

در این آزمایش، الگوریتم پیشنهادی را با شرایط زیر اجرا کرده‌ایم:

۱. معیار انتخاب نقطه‌ی شکست = شاخص جینی
۲. m = nVariable به معنی آن که پارامتر m برابر با تعداد کل ویژگی‌های موجود است. باتوجه‌به دو مجموعه‌داده‌ی در دسترس، با به‌کارگیری مجموعه‌داده‌ی اول این پارامتر مقدار ۱۴ و با به‌کارگیری مجموعه‌داده‌ی دوم، مقدار ۲۰ را خواهد داشت.
۳. تعداد انشعابات = متغیر بوده و از مقدار ۱ با گام افزایشی ۱۰ و تا مقدار ۲۰۰ مقداردهی خواهد شد.

معیارهای ارزیابی الگوریتم‌های دسته‌بندی که در بخش قبل به آن‌ها اشاره شد، هم در فرایند اجرای الگوریتم محاسبه و در جدول ۲ نشان داده شده است.

جدول (۲) معیارهای ارزیابی محاسبه شده بر حسب تعداد انشعابات در مدل پیشنهادی بر مجموعه‌داده‌ی استرالیایی با ۱۴ ویژگی

تعداد انشعابات	Accuracy	Sensitivity	Precision	F-measure
۱	0.84	0.82	0.86	0.84
۱۰	0.9	0.92	0.9	0.91
۲۰	0.89	0.9	0.9	0.9
۳۰	0.91	0.91	0.92	0.92
۴۰	0.92	0.93	0.92	0.92

0.93	0.93	0.92	0.92	۵۰
0.92	0.93	0.92	0.92	۶۰
0.94	0.95	0.94	0.94	۷۰
0.93	0.94	0.92	0.92	۸۰
0.94	0.94	0.93	0.93	۹۰
0.94	0.94	0.93	0.93	۱۰۰
0.94	0.94	0.93	0.93	۱۱۰
0.94	0.95	0.93	0.94	۱۲۰
0.94	0.94	0.94	0.93	۱۳۰
0.94	0.94	0.94	0.93	۱۴۰
0.94	0.94	0.94	0.93	۱۵۰
0.93	0.94	0.92	0.92	۱۶۰
0.94	0.94	0.94	0.94	۱۷۰
0.94	0.95	0.94	0.94	۱۸۰
0.94	0.94	0.94	0.93	۱۹۰
0.94	0.95	0.93	0.93	۲۰۰

جدول (۳) معیارهای ارزیابی محاسبه شده بر حسب تعداد انشعابات در مدل پیشنهادی بر مجموعه‌داده‌ی آلمانی با ۲۰ ویژگی

F-measure	Precision	Sensitivity	Accuracy	تعداد انشعابات
0.86	0.81	0.92	0.8	۱
0.85	0.89	0.81	0.8	۱۰
0.87	0.88	0.86	0.83	۲۰
0.89	0.89	0.88	0.85	۳۰
0.9	0.91	0.89	0.86	۴۰
0.91	0.93	0.89	0.88	۵۰
0.92	0.92	0.91	0.89	۶۰

0.92	0.91	0.93	0.89	۷۰
0.92	0.92	0.92	0.89	۸۰
0.93	0.93	0.93	0.91	۹۰
0.92	0.93	0.92	0.9	۱۰۰
0.93	0.91	0.94	0.9	۱۱۰
0.93	0.91	0.94	0.9	۱۲۰
0.93	0.92	0.95	0.91	۱۳۰
0.93	0.91	0.95	0.9	۱۴۰
0.94	0.93	0.95	0.91	۱۵۰
0.94	0.94	0.94	0.92	۱۶۰
0.94	0.94	0.94	0.92	۱۷۰
0.94	0.93	0.95	0.91	۱۸۰
0.94	0.94	0.94	0.92	۱۹۰
0.94	0.94	0.94	0.92	۲۰۰

همان‌گونه که مشخص است با افزایش تعداد انشعاب‌ها معیارهای ارزیابی نیز روند صعودی دارند. پس قدرت الگوریتم در شناسایی نیز افزایش یافته است. باتوجه‌به جدول ۱ که مربوط به انجام آزمایش بر روی مجموعه‌داده‌ی استرالیای با ۱۴ ویژگی است، مشاهده می‌کنیم زمانی که تعداد انشعاب‌ها به ۹۰ می‌رسد تقریباً رفتار مدل نیز قابل‌پیش‌بینی بوده و تغییرات چندانی در معیارها مشاهده نمی‌شود. همچنین بهترین نتایج با تعداد انشعاب ۱۲۰، به دست آمده است. علاوه بر این، باتوجه‌به نتایج به دست آمده در جدول ۲ که مربوط به انجام آزمایش بر روی مجموعه‌داده‌ی آلمانی با ۲۰ ویژگی است، همین نکته را دریافتیم که با افزایش تعداد انشعابات در مدل دقت تشخیص الگوریتم افزایش می‌یابد. نکته‌ی قابل‌توجه دیگر آن است با افزایش تعداد ویژگی‌های مسئله به تعداد انشعاب‌ها بیش‌تری نیاز است تا به‌دقت مطلوب دست یابیم. چنان‌که مشاهده می‌کنیم زمانی که تعداد انشعاب به ۱۶۰ می‌رسد، رفتار مدل تقریباً ثابت می‌شود درحالی‌که در مجموعه‌داده‌ی قبلی که ۱۴ ویژگی داشت، با داشتن ۹۰ انشعاب به کارایی مطلوب دست می‌یابیم.

درجه کوادراتیک

یکی از پارامترهای مهم در تعیین بهترین مدل در فضای ویژگی، درجه کرنل کوادراتیک در مدل پیشنهادی است. هر چه تعداد ویژگی‌ها و درجه کرنل افزایش می‌یابد، انتخاب ویژگی مناسب‌تر و در نتیجه دقت تشخیص بهبود می‌یابد.

برای مشاهده‌ی تأثیر این پارامتر در کارایی الگوریتم و انتخاب درجه مناسب آن در مدل، آزمایش زیر انجام شده است:

آزمایش ۲: تأثیر پارامتر درجه کرنل کوادراتیک

در این آزمایش، الگوریتم پیشنهادی را با شرایط زیر اجرا کرده‌ایم:

۱. معیار انتخاب نقطه‌ی شکست = شاخص جینی

۲. m = nVariable به معنی آن که پارامتر m برابر با تعداد کل ویژگی‌های موجود است. باتوجه‌به دو مجموعه‌داده‌ی در دسترس، با به‌کارگیری مجموعه‌داده‌ی اول این پارامتر مقدار ۱۴ و با به‌کارگیری مجموعه‌داده‌ی دوم، مقدار ۲۰ را خواهد داشت.

۳. درجه کرنل = متغیر بوده و از مقدار ۱ تا مقدار ۳ مقداردهی خواهد شد.

معیارهای ارزیابی الگوریتم‌های دسته‌بندی که در بخش قبل به آن‌ها اشاره شد، هم در فرایند اجرای الگوریتم محاسبه و در جدول ۴ نشان داده شده است.

جدول (۴) معیارهای ارزیابی محاسبه شده بر حسب کرنل کوادراتیک در مدل پیشنهادی بر مجموعه‌داده‌ی استرالیایی با ۱۴ ویژگی

کرنل کوادراتیک بر روی فضای ویژگی انشعاب‌ها	Accuracy	Sensitivity	Precision	F-measure
تک‌هسته‌ای	0.82	0.81	0.82	0.83
دو هسته‌ای	0.91	0.88	0.89	0.90
سه هسته‌ای	0.94	0.95	0.95	0.94

جدول (۵) معیارهای ارزیابی محاسبه شده بر حسب کرنل کوادراتیک در مدل پیشنهادی بر مجموعه‌داده‌ی آلمانی با ۲۰ ویژگی

کرنل کوادراتیک بر روی فضای ویژگی انشعاب‌ها	Accuracy	Sensitivity	Precision	F-measure
تک‌هسته‌ای	0.84	0.82	0.84	0.82
دو هسته‌ای	0.93	0.91	0.92	0.92
سه هسته‌ای	0.96	0.96	0.96	0.95

همان‌گونه که مشخص است با افزایش تعداد ابعاد کرنل معیارهای ارزیابی نیز روند صعودی دارند. پس قدرت الگوریتم در شناسایی نیز افزایش یافته است. باتوجه‌به جداول فوق که مربوط به انجام آزمایش بر روی مجموعه‌داده‌ی استرالیای با ۱۴ ویژگی و مربوط به انجام آزمایش بر روی مجموعه‌داده‌ی آلمانی با ۲۰ ویژگی است، دریافتیم که با افزایش ابعاد کرنل از خطی به درجه سه در مدل دقت تشخیص الگوریتم افزایش می‌یابد. نکته‌ی قابل‌توجه دیگر آن است با افزایش تعداد ویژگی‌های مسئله استفاده از کوادراتیک نتایج مطلوب‌تری به همراه داشته است.

پارامتر m

یکی دیگر از پارامترهایی که در مدل پیشنهادی قابل‌تنظیم می‌باشد، با نام m شناخته می‌شود. یادآور می‌شویم که این پارامتر به معنی تعداد ویژگی‌هایی است که در هر گره به طور تصادفی از مجموعه‌ی کل ویژگی‌های مسئله استخراج شده و با استفاده از این زیرمجموعه‌ی انتخاب شده بهترین ویژگی برای شکست انتخاب می‌شود. در طول زمان ساخت مدل، این پارامتر برای تمامی انشعابات ثابت در نظر گرفته می‌شود و مقادیر معمولی که برای آن انتخاب می‌گردد،Sqrt(nVariable) و یا Log(nVariable) خواهد بود. برای مشاهده‌ی تأثیر این پارامتر و انتخاب بهترین مقدار ممکن برای آن، آزمایش زیر را اجرا کرده‌ایم:

آزمایش ۳: تأثیر پارامتر m

در این آزمایش مدل پیشنهادی را با شرایط زیر اجرا کرده‌ایم:

۱. تعداد انشعاب‌ها = ۱۲۰

۲. معیار تعیین ویژگی شکست = شاخص جینی

۳. پارامتر m متغیر بوده و سه مقدار مجزا به خود می‌گیرد. ۱. جذر تعداد کل ویژگی‌ها در هر مجموعه داده. ۲. لگاریتم تعداد کل ویژگی‌ها در هر مجموعه داده. ۳. تعداد کل ویژگی‌ها در هر مجموعه داده.

در جداول ۵ و ۶ نتایج ارزیابی الگوریتم در این سه حالت مختلف و بر روی دو مجموعه‌داده‌ی موجود، داده نشان داده شده است.

جدول (۵) معیارهای ارزیابی مدل پیشنهادی بر حسب مقادیر مختلف پارامتر m، مجموعه‌داده‌ی استرالیایی با ۱۴ ویژگی

F-measure	Sensitivity	Accuracy	پارامتر m
0.93	0.93	0.93	Sqrt(nVariable)
0.93	0.94	0.92	Log (nVariable)
0.93	0.93	0.93	nVariable

جدول (۶) معیارهای ارزیابی مدل پیشنهادی بر حسب مقادیر مختلف پارامتر m، مجموعه‌داده‌ی آلمانی با ۲۰ ویژگی

F-measure	Sensitivity	Accuracy	پارامتر m
0.92	0.93	0.89	Sqrt(nVariable)
0.92	0.92	0.89	Log (nVariable)
0.93	0.93	0.9	nVariable

باتوجه‌به نتایج فوق می‌توان به این نکته پی برد که در مسئله‌ی حاضر و با مجموعه‌داده‌های در دسترس کاهش تعداد ویژگی‌ها به مقادیر sqrt(nVariable) و log(nVariable) نتایج امیدبخشی را در پی ندارد.

معیار تعیین نقطه‌ی شکست

یکی از مهم‌ترین پارامترهای موجود در الگوریتم پیشنهادی و کلیه‌ی روش‌هایی که بر مبنای درخت تصمیم عمل می‌کنند، انتخاب معیاری مناسب برای تعیین نقطه‌ی شکست در زمان ساخت مدل خواهد بود. ازآنجایی‌که هسته‌ی اصلی مدل پیشنهادی، درختان تصمیم موجود در آن است، پس انتخاب بهترین معیار برای تعیین نقطه یا همان ویژگی برای بخش‌بندی داده‌ها نکته‌ی مهمی محسوب می‌شود. همان‌طور که ذکر شد، معیارهای مختلفی برای این کار وجود دارد که به‌تفصیل توصیف شدند.

برای مشاهده‌ی تأثیر این پارامتر و انتخاب بهترین معیار انتخاب نقطه‌ی شکست، آزمایش سوم به شکل زیر انجام شده است:

آزمایش ٤: تأثیر پارامتر تعیین ویژگی شکست

در این آزمایش مدل پیشنهادی را با شرایط زیر اجرا کرده‌ایم:

- تعداد انشعابات = ۱۲۰

۴. m = nVariable به معنی آن که پارامتر m برابر با تعداد کل ویژگی‌های موجود است. باتوجه‌به دو مجموعه‌داده‌ی در دسترس، با به‌کارگیری مجموعه‌داده‌ی اول این پارامتر مقدار ۱۴ و با به‌کارگیری مجموعه‌داده‌ی دوم، مقدار ۲۰ را خواهد داشت.

- معیار انتخاب نقطه‌ی شکست = متغیر بوده و هر بار یکی از معیارهای بهره‌ی اطلاعاتی، نسبت بهره و شاخص جینی انتخاب و آزمایش می‌شود.

در جداول ۷ تا ۱۱ نتایج ارزیابی الگوریتم در این سه حالت مختلف و بر روی دو مجموعه‌داده‌ی موجود، داده نشان داده شده است.

جدول (۷) معیارهای ارزیابی مدل پیشنهادی بر حسب معیارهای مختلف تعیین ویژگی بخش‌بندی داده‌ها، مجموعه‌داده‌ی استرالیایی با ۱۴ ویژگی

F-measure	Sensitivity	Accuracy	
0.72	0.6	0.9	بهره اطلاعاتی
0.55	0.4	0.72	نسبت بهره
0.93	0.95	0.93	شاخص جینی

جدول (۸) معیارهای ارزیابی مدل پیشنهادی بر حسب معیارهای مختلف تعیین ویژگی بخش‌بندی داده‌ها، مجموعه‌داده‌ی آلمانی با ۲۰ ویژگی

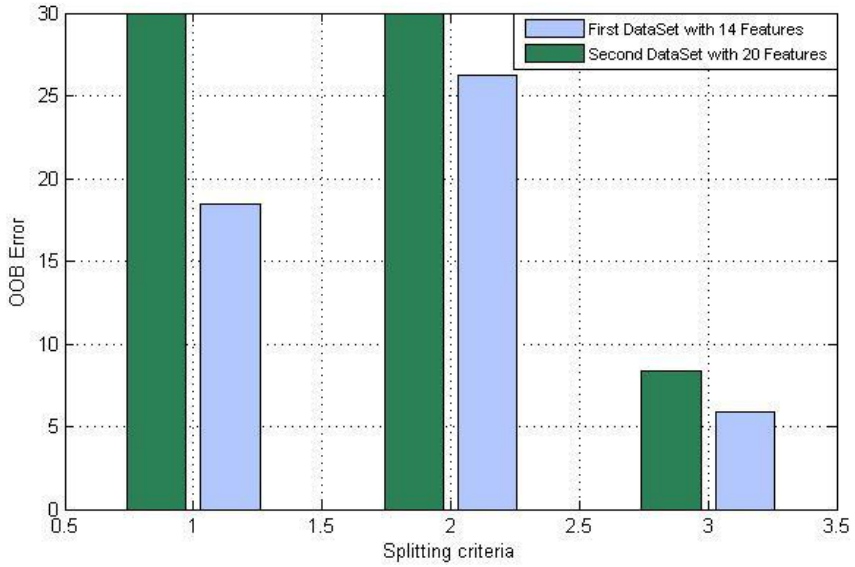
F-measure	Sensitivity	Accuracy	
0.85	0.86	0.75	بهره اطلاعاتی
0.76	0.8	0.7	نسبت بهره
0.95	0.93	0.96	شاخص جینی

جدول (۹) معیارهای ارزیابی مدل پیشنهادی برحسب معیارهای مختلف تعیین ویژگی بخش‌بندی داده‌ها، مجموعه‌داده‌ی استرالیایی با ۱۴ ویژگی

Precision	خطای OOB	
0.93	18.4%	بهره اطلاعاتی
0.93	26.2%	نسبت بهره
0.95	4.6%	شاخص جینی

جدول (۱۰) معیارهای ارزیابی مدل پیشنهادی برحسب معیارهای مختلف تعیین ویژگی بخش‌بندی داده‌ها، مجموعه‌داده‌ی آلمانی با ۲۰ ویژگی

Precision	خطای OOB	
0.9	30%	بهره اطلاعاتی
0.88	30%	نسبت بهره
0.95	4.2%	شاخص جینی



شکل (۱) خطای OOBبرحسب معیارهای مختلف تعیین ویژگی بخش‌بندی داده‌ها

باتوجه‌به نتایج به دست آمده از این آزمایش در مسئله‌ی حاضر و باتوجه‌به مجموعه‌داده‌های در دسترس، مشاهده می‌شود که استفاده از معیار شاخص جینی میزان دقت بیش‌تر شده است و نتایج بهتری را در سایر معیارهای ارزیابی نسبت به دو معیار دیگر فراهم آورده است. همچنین مشاهده می‌شود که با استفاده از معیار شاخص جینی میزان صحت بیش‌تر و خطای OOB کمتر شده است.

محدودیت‌های روش پیشنهادی

از جمله محدودیت‌های موجود افزایش تعداد تراکنش‌های بانکی و مواجهه با حجم بسیار زیادی از اطلاعات به‌روز شده در هر لحظه است.

محدودیت دیگر این است که چون عملیات بانکی نیازمند پاسخ بلادرنگ می‌باشند، زمان بسیار محدودی برای بررسی تراکنش و مسدودکردن آن در صورت مشکوک بودن می‌باشد.

یکی از محدودیت‌های دیگر در این زمینه اطلاعات غیرساختاریافته است. به همین دلیل روش شناسایی تقلب باید قادر باشد این اطلاعات را به‌درستی سازماندهی کند تا بتواند تحلیل و تشخیص درستی ارائه دهد.

بحث و مقایسه

در بخش قبل با انجام آزمایش‌های مختلف سعی در یافتن بهترین مقادیر برای پارامترهای مختلف این الگوریتم داشتیم؛ بنابراین باتوجه‌به این آزمایش اول دریافتیم که مناسب‌ترین تعداد انشعاب برای مدلی که داده‌های تست را دسته‌بندی کند، در هر دو مجموعه‌داده‌ی در دسترس، ۱۸۰ بوده است. این ادعا باتوجه‌به نتایج موجود در جداول ۱ و ۲ ثابت می‌شود.

با انجام آزمایش دوم، سعی در یافتن بهترین مقدار برای پارامتر m در این الگوریتم داشتیم. باتوجه‌به نتایج به دست آمده در جداول ۳ و ۴، اگر این پارامتر را به تعداد ویژگی‌های موجود در مجموعه‌داده مقداردهی کنیم، بهترین شرایط در معیارهای ارزیابی حاصل می‌شود. در واقع باتوجه‌به مجموعه‌داده‌های در دسترس و محدود بودن تعداد ویژگی در آن‌ها، کاهش مقدار این پارامتر به جذر تعداد کل ویژگی‌ها و یا لگاریتم آن، باعث کاهش کارایی الگوریتم شده و در واقع مدل حاصل به‌خوبی آموزش داده نشده و با تمامی فضای مسئله منطبق نمی‌گردد.

در نهایت، هدف از انجام آزمایش سوم تعیین بهترین معیار انتخاب ویژگی شکست در گره‌ها بود. باتوجه‌به نتایج نشان داده شده در جداول ۴ و ۵ دریافتیم که استفاده از معیار شاخص جینی در هر دو مجموعه‌داده عملکرد مناسبی دارد.

در این بخش برای نمایش قدرت و کارایی مدل پیشنهادی، این الگوریتم را با الگوریتم درخت تصمیم و ماشین بردار پشتیبان و شبکه عصبی و جنگل تصادفی معمولی که از معروف‌ترین الگوریتم‌های داده‌کاوی محسوب می‌شوند و همچنین روش ارائه شده در مقاله اوش و همکاران ارزیابی می‌کنیم. روش ارائه شده در این مطالعه یک شبکه عصبی را برای شناسایی تقلب در کارت‌های اعتباری و آزمایش آن بر روی مجموعه‌داده‌ی آلمانی با ۲۰ ویژگی که ما در این پروژه از آن استفاده کرده‌ایم، به کار گرفته است. در این تحقیق، شبکه‌ی عصبی با الگوریتم شبیه‌سازی ذوب فلزات آموزش داده شده است. به معنی آن که اوزان موجود در شبکه با این الگوریتم آموزش داده شده‌اند. در واقع به‌جای استفاده از روش آموزش گرادیان نزولی که تاکنون بسیار مورداستفاده قرار گرفته است و در اکثر مواقع در کمینه‌ی محلی گیر می‌افتد، از این الگوریتم هیوریستیک که توان فرار از کمینه-های محلی را دارد، بهره گرفته و سعی در افزایش دقت شبکه عصبی حاصل داشته است. میزان دقت (Accuracy) به دست آمده با به‌کارگیری این روش،89.6% گزارش شده است. با اجرای الگوریتم پیشنهادی با شرایط موجود در جدول 11، به دقت تشخیص ۹۶٪ در مجموعه‌داده‌ی آلمانی با ۲۰ ویژگی و دقت تشخیص 95% در مجموعه‌داده‌ی استرالیایی با ۱۴ ویژگی، دست‌یافته‌ایم.

جدول (۱۱) پارامترهای تنظیم شده برای ساخت مدل و ارزیابی آن

پارامتر	مقدار
تعداد انشعابات	180
پارامتر m	nVariable
معیار تعیین ویژگی شکست	شاخص جینی

همچنین برای ساخت مدل، در هر دو مجموعه‌داده، از ۷۰ درصد داده‌ها برای آموزش و از ۳۰ درصد باقیمانده برای تست آن استفاده کرده‌ایم. معیار انتخاب نقطه‌ی شکست در انشعاب نیز شاخص جینی انتخاب شده است. چراکه مشاهده کردیم این معیار نسبت به بقیه نتایج بهتری در پی دارد.

برای ساخت مدل ماشین بردار پشتیبان نیز از ۷۰ درصد داده‌ها برای آموزش و از ۳۰ درصد دیگر برای تست آن استفاده کردیم. همچنین از تابع خطی به‌عنوان تابع kernel در این مدل استفاده کرده‌ایم. سپس هر دو الگوریتم را بر روی هر دو مجموعه داده اجرا کرده، معیارهای ارزیابی را برای هر دو مدل محاسبه و در جداول ۱۲ تا ۱۵ با الگوریتم پیشنهادی مقایسه شده است. همان‌طور که مشخص است مدل پیشنهادی این پژوهش در مقایسه با این روش‌ها کاملاً بهتر عمل کرده و امید بخش بوده و استفاده از آن دقت عملکرد بهتری را فراهم آورده است.

جدول (۱۲) معیارهای ارزیابی الگوریتم پیشنهادی و درخت تصمیم، ماشین بردار پشتیبان، شبکه عصبی و جنگل تصادفی معمولی مجموعه داده استرالیایی با ۱۴ ویژگی

الگوریتم	Accuracy	Sensitivity	F-measure
مدل پیشنهادی	0.95	0.95	0.95
درخت تصمیم	0.9	0.9	0.91
ماشین بردار پشتیبان	0.75	0.84	0.81
شبکه عصبی	0.91	0.92	0.91
جنگل تصادفی معمولی	0.94	0.93	0.94

جدول (۱۳) معیارهای ارزیابی الگوریتم پیشنهادی و درخت تصمیم، ماشین بردار پشتیبان، شبکه عصبی و جنگل تصادفی معمولی مجموعه داده آلمانی با ۲۰ ویژگی

الگوریتم	Accuracy	Sensitivity	F-measure
مدل پیشنهادی	0.96	0.96	0.95
درخت تصمیم	0.77	0.75	0.82
ماشین بردار پشتیبان	0.86	0.9	0.91
شبکه عصبی	0.87	0.9	0.89
جنگل تصادفی معمولی	0.92	0.92	0.93

جدول (۱۴) معیارهای ارزیابی الگوریتم پیشنهادی و درخت تصمیم، ماشین بردار پشتیبان، شبکه عصبی و جنگل تصادفی معمولی مجموعه داده استرالیایی با ۱۴ ویژگی

الگوریتم	خطای OOB	Precision
مدل پیشنهادی	4.2%	0.96
درخت تصمیم	-	0.92
ماشین بردار پشتیبان	-	0.71
شبکه عصبی	-	0.92
جنگل تصادفی معمولی	8.4%	0.94

جدول (۱۶) معیارهای ارزیابی الگوریتم پیشنهادی و درخت تصمیم، ماشین بردار پشتیبان، شبکه عصبی و جنگل تصادفی معمولی مجموعه داده آلمانی با ۲۰ ویژگی

الگوریتم	خطای OOB	Precision
مدل پیشنهادی	2.8%	0.96
درخت تصمیم	-	0.89
ماشین بردار پشتیبان	-	0.83

شبکه عصبی	-	0.91
جنگل تصادفی معمولی	12.6%	0.92

نتیجه‌گیری

امروزه استفاده از کارت‌های بانکی در سطوح گسترده‌ای از تعاملات تجاری مطرح است. هرچند این تحولات گامی بزرگ در جهت کارایی و سهولت دسترسی است، معایبی نیز به همراه دارد که مهم‌ترین آن آسیب‌پذیری نسبت به فعالیت‌های متقلبانه است؛ به همین دلیل پژوهشگران همواره دنبال ارائه‌ی روشی جدید و مؤثر و کارا برای شناسایی به هنگام و یا پیشگیری از وقوع تقلب در کارت‌های بانکی هستند.

در این مقاله برای یافتن بهترین مقادیر برای پارامترهای مورد تنظیم در این الگوریتم آزمایشاتی طراحی و اجرا کردیم. پارامترهای آزاد در این الگوریتم عبارت‌اند از تعداد انشعابات، درجه کوادراتیک، پارامتر m که دامنه‌ی انتخاب ویژگی شکست را معین می‌کند و معیار تعیین ویژگی شکست. با انجام چهار آزمایش مقادیر این پارامترها برآورد شده و نتایج حاصل از آن‌ها ارائه گردید. سپس الگوریتم پیشنهادی را ارزیابی کردیم. با انجام آزمایش نهایی و ارزیابی الگوریتم دریافتیم که مدل ارائه شده در این پژوهش با قدرت بیش‌تری عمل کرده و کاملاً رضایت‌بخش و نتایج خوبی در پی داشته است. حاصل پژوهش انجام شده مدلی برای شناسایی متقلبین در سیستم‌های بانکداری آنلاین بر مبنای تراکنش‌های کارت‌های اعتباری بوده است که عملکرد آن در دسته‌بندی تراکنش‌ها و متقلبین مناسب به نظر می‌رسد.

پیشنهادهای آینده

پژوهش حاضر به طور خاص بر روی تشخیص متقلبین در سیستم‌های بانکداری آنلاین معطوف شده است. برای ارزیابی روش ارائه شده در این مقاله، می‌توان این مدل را در سایر زمینه‌های مشابه تحقیقاتی مثل شناسایی تقلب در بازار بورس و خریدوفروش سهام که در آن نیز تعاملات مهمی انجام می‌شود و به دلیل طرف‌داران زیادی که سالانه به خود جذب می‌کند می‌تواند بسیار موردتوجه افراد متقلب برای فریب افراد تازه‌وارد، قرار گیرد. می‌توان در پژوهش‌های بعدی این مدل را در زمینه‌های دیگر به کار گرفته و آن را ارزیابی کرد.

فهرست مراجع

بنائی، هادی، خوش‌نیت، حسام. (۱۳۹۶). نقش و کاربرد هوش عملیاتی و داده‌کاوی در کشف تقلب برخط. ششمین همایش ملی تجارت و اقتصاد الکترونیک. همایش تخصصی امنیت و اعتماد.

حاتمی‌راد، علی، شهریاری، حمیدرضا. (۱۳۹۷). روش‌ها و راهکارهای شناسایی تقلب در بانک‌داری الکترونیک. فصل‌نامه تازه‌های اقتصاد، سال نهم، شماره ۱۳۴، صص ۲۱۹ تا ۲۲۸.

قلی پور سلیمانی، علی، ایمانی، سهیلا. (۱۴۰۰) سیر تکنولوژی در بانکداری.دو ماهنامه مدیریت، شماره. ۱۵۹ صص ۲۲ تا ۲۵.

وثوق، ملیحه، تقوی‌فرد، محمدتقی و البرزی، محمود. (۱۳۹۸). شناسایی تقلب در کارت‌های بانکی با استفاده از شبکه‌های عصبی مصنوعی. فصل‌نامه علمی-پژوهشی مدیریت فناوری اطلاعات دانشگاه تهران، دوره ۶، شماره ۴، صص ۷۲۱-۷۴۶.

Ata, H. A., & Seyrek, I. H. (2009). THE USE OF DATA MINING TECHNIQUES IN DETECTING FRAUDULENT FINANCIAL STATEMENTS: AN APPLICATION ON MANUFACTURING FIRMS. *Suleyman Demirel University Journal of Faculty of Economics & Administrative Sciences*, *14*(2).

Bahnsen, A. C., Aouada, D., & Ottersten, B. (2015). Example-dependent cost-sensitive decision trees. *Expert Systems with Applications*, *42*(19), 6609-6619.

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

Bansal, M., & Sharma, D. (2021). A novel multi-view clustering approach via proximity-based factorization targeting structural maintenance and sparsity challenges for text and image categorization. *Information Processing & Management*, *58*(4), 102546.

Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision support systems*, *50*(3), 602-613.

Bose, I., & Mahapatra, R. K. (2001). Business data mining—a machine learning perspective. *Information & management*, *39*(3), 211-225.

Breiman, L. (2011). Random forests. *Machine learning*, *45*, 5-32.

Carta, S., Fenu, G., Recupero, D. R., & Saia, R. (2019). Fraud detection for E-commerce transactions by employing a prudential Multiple Consensus model. *Journal of Information Security and Applications*, *46*, 13-22.

Chandra, V., & Singh, P. (2014). Fuzzy Based High Blood Pressure Diagnosis. *International Journal of Advanced Research in Computer Science & Technology (IJARCST)*, *2*(2), 2347–8446.

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.

Dreżewski, R., Sepielak, J., & Filipkowski, W. (2015). The application of social network analysis algorithms in a system supporting money laundering detection. *Information Sciences*, *295*, 18-32.

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

Fang, W., Li, X., Zhou, P., Yan, J., Jiang, D., & Zhou, T. (2021). Deep learning anti-fraud model for internet loan: where we are going. *IEEE Access*, *9*, 9777-9784.

Hirshman, J., Huang, Y., & Macke, S. (2013). Unsupervised approaches to detecting anomalous behavior in the bitcoin transaction network. *Technical report*, *Stanford University*.

JYeonkook J. Kim, Bok Baik b, Sungzoon Cho, "Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning", Expert Systems With Applications, Vol. 62, Pages 32–43, (2019).

Moreira, M. Â. L., Junior, C. D. S. R., de Lima Silva, D. F., de Castro Junior, M. A. P., de Araújo Costa, I. P., Gomes, C. F. S., & dos Santos, M. (2022). Exploratory analysis and implementation of machine learning techniques for predictive assessment of fraud in banking systems. *Procedia Computer Science*, *214*, 117-124.

Nazeer, I., Prasad, K. D. V., Bahadur, P., Bapat, V., & MJ, K. (2023). Synchronization of AI and Deep Learning for Credit Card Fraud Detection. *International Journal of Intelligent Systems and Applications in Engineering*, *11*(5s), 52-59.

Patidar, R., & Sharma, L. (2011). Credit card fraud detection using neural network. *International Journal of Soft Computing and Engineering (IJSCE)*, *1*(32-38).

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*.

Salchenberger, L. M., Cinar, E. M., & Lash, N. A. (1992). Neural networks: A new tool for predicting thrift failures. *Decision Sciences*, *23*(4), 899-916.

Shen, A., Tong, R., & Deng, Y. (2007, June). Application of classification models on credit card fraud detection. In *2007 International conference on service systems and service management* (pp. 1-4). IEEE.

Van Vlasselaer, V., Bravo, C., Caelen, O., Eliassi-Rad, T., Akoglu, L., Snoeck, M., & Baesens, B. (2015). APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions. *Decision Support Systems*, *75*, 38-48.

Wang, X., Wang, X., Wilkes, M., Wang, X., Wang, X., & Wilkes, M. (2021). A k-nearest neighbour spectral clustering-based outlier detection technique. *New Developments in Unsupervised Outlier Detection: Algorithms and Applications*, 147-172.