سومین کنفرانس ملی مباحث نوین در کامپیوتر و فناوری اطلاعات
3ed National Conference on Advanced Topics in Computer and Information Technology
بیست و هشتم آذر ماه ۱۳۹۸

بکارگیری تکنیکهای داده کاوی در تشخیص و پیش بینی کلاهبرداری بانکی

زهرا رحیمی(۱) ، محمدامین شایگان(۲،*)
(۱) گروه مهندسی کامپیوتر، واحد شیراز، دانشگاه آزاد اسلامی، شیراز، ایران
Zahra70rahimi@gmail.com
(۲) گروه مهندسی کامپیوتر ، واحد شیراز ، دانشگاه آزاد اسلامی، شیراز، ایران
* نویسنده مسئول : Shayegan@iaushiraz.ac.ir

خلاصه: با گسترش روز افزون استفاده از سامانه های نوین بانکی و افزایش تعداد عملیات بانکی، سوء استفاده های مالی و تقلب در این عملیات بیش از پیش گسترش پیدا کرده است. اینگونه سوء استفاده ها علاوه بر اتلاف منابع مالی، باعث کاهش اعتماد مشتریان به استفاده از سامانه های نوین بانکی و در نتیجه کاهش اثر بخشی این سامانه ها در مدیریت بهینه ی سرمایه و تراکنش های مالی می شود. در این پژوهش جهت کشف تقلب بانکی بر روی مجموعه داده های بانکی ، از ترکیب الگوریتم های داده کاوی استفاده شده است. برای انجام کار در ابتدا، خوشه بندی رکوردهای داده ای موجود در مجموعه داده ها صورت گرفته است و به دنبال آن، تشخیص تراکنش های بانکی شبهه دار، در زمان انجام تراکنش تشخیص داده می شود. نتایج حاصل نشان می دهد که روش پیشنهادی دارای میزان دقت بالاتری نسبت به الگوریتم های داده کاوی دیگر همچون درخت تصمیم J48 و جنگلهای تصادفی دارد.

کلمات کلیدی: داده کاوی، کشف تقلب، تراکنش های بانکی ، درخت تصمیم ، جنگلهای تصادفی .

۱ – مقدمه

طی دو دهه اخیر، اهمیت تجارت الکترونیک[1] به طور چشمگیری افزایش یافته و همچنان رو به افزایش است. امروزه استفاده از تجارت الکترونیک و سرویس های ارتباطی و اطلاعاتی، برای دسترسی بهتر و بیشتر مشتریان به طور فزاینده ای رواج یافته است. بسیاری از شرکت ها و مؤسسه ها، بخشی از کسب وکار خود را به سمت خدمات برخط[2] سوق داده اند. صنعت بانکداری نیز از این فناوری ها بی بهره نمانده و با ایجاد خدمات الکترونیکی و نظام های پرداخت[3]، موجب کاهش تعاملات فیزیکی در محیط اداری بانکها شده و استفاده از خدمات بانکها را به سمت منازل و محیط کار افراد سوق داده است. یکی از خدمات بانکهای ایرانی در سالهای اخیر، که با استقبال زیاد مشتریان بانکها روبه رو شد، استفاده از کارتهای بانکی در سطوح گسترده ای از تعاملات تجاری است.

هرچند تحولات یاد شده گامی بزرگ در جهت کارایی، سهولت دسترسی و سودآوری است، معایبی نیز دارد که مهم ترین آن، آسیب پذیری نسبت به تهدیدهاست. چرا که بسیاری از تخلف های نظام بانکی و فعالیت های متقلبانه، به سیستم های بانکداری الکترونیکی بازمی گردد. کارتهای بانکی، یکی از دلایل عمده رشد بانکداری الکترونیک، اکنون به پرا ستفاده ترین ابزار بانکداری تبدیل شده است، لذا بخش عمده ای از فعالیت های متقلبانه، معطوف به تراکنش با این کارتهاست.

اشخاص حقوقی، حقیقی و همچنین بانکها، سالانه مبالغ هنگفتی را به واسطه تقلب و متقلبانی از دست می دهند که دائم به دنبال یافتن راههای جدیدی برای اقدامات غیرقانونی با استفاده از این کارتها هستند. تقلب در استفاده از کارت اعتباری در سراسر جهان به یک مشکل بزرگ تبدیل شده است [۱]. در سال ۲۰۱۲، با افزایش ۱۴/۸٪ در مقایسه با سال ۲۰۱۱، کل میزان تقلب برابر با ۱,۳۳ میلیارد یورو بوده است. علاوه بر این، پرداختها در کانالهای ارتباطی جدید (تلفن همراه، اینترنت) برابر با ۶۰٪ کلاهبرداری بود. در حالیکه در سال

[1] e-Commerce
[2] Online
[3] Payment Systems

۲۰۰۸ این آمار ۴۶٪ بوده است. این مسئله چالشهای جدیدی را بوجود می آورد. لذا برای جلوگیری از کلاهبرداری با توجه به تغییر استراتژی نیاز به ارائه رویکردهای جدیدی است [۱].

تاکنون در سیستم بانکی کشور ایران، سازوکار و برنامه جامعی برای شناسایی و جلوگیری از تقلب های مربوط به تراکنش های مبتنی بر کارت وجود نداشته است و به دلیل نبودن سیستمی مناسب، تقلب های زیادی ناشناخته باقی مانده اند. در سایر کشورها نیز به دلیل گستردگی استفاده از کارتهای اعتباری، پژوهش های انجام گرفته به طور عمده بر این کارتها تمرکز کرده اند؛ درحالی که استفاده از این نوع کارتها در کشور ایران هنوز رواج پیدا نکرده و کمابیش همه تراکنش های به وسیله کارتهای نقدی (از پیش پرداخت شده[4]) صورت می گیرد. همچنین، با توجه به ملاحظات امنیتی، نتایج مطالعات صورت گرفته به طور کامل منتشر نمی شوند و نمی توان از آنها بهره ای برد. بنابراین بهره گیری از مدلهای طراحی شده در ادبیات تحقیق سایر کشورها چندان مقدور نیست.

با توجه به حجم گسترده تراکنش های روزانه بانکی و نیاز به تشخیص به موقع تقلب ها و جلوگیری از وقوع آنها، در عمل شناسایی دستی این تقلب ها امکان پذیر نیست و مستلزم صرف زمان و نیروی انسانی بسیاری خواهد بود [۲]. بنابراین با توجه به نبود سازوکاری برای شناسایی تقلب در کارتهای سیستم بانکی کشور، مسئله اصلی این پژوهش، ایجاد چارچوبی برای شناسایی تقلب در کارتهای بانکی، هنگام تراکنش یا به فاصله کوتاهی پس از آن است. بدین منظور از ابزار داده کاوی استفاده شده و بانک صادرات، به عنوان یک مطالعه موردی مورد بررسی قرار گرفته است. استفاده از کارت توسط مشتری مشخص، معمولاً از الگوهای مشخصی تبعیت می کند که شبکه عصبی با استفاده از بازشناسی الگو می تواند شناسایی این الگوها و تقلب های مربوط به آن را امکان پذیر کند [۳]. لذا استفاده از این تکنیک در این پژوهش مدنظر خواهد بود.

یکی از حساس ترین، وظایف بانکها، نظارت بر صحت و سلامت تراکنش های انجام گرفته روی حسابها، به منظور حفظ امنیت مشتریان بانکها و همچنین خود بانکها است. از اینرو ایجاد سیستمی به منظور شناسایی تقلب در کارتهای بانکی، با استفاده از داده کاوی که ناظر بر عملکرد نظام های پرداخت باشد، ضروری به نظر می رسد. یکی از اصلی ترین زیرساختهای ایجاد چنین سیستمی، تدوین روشی مناسب برای شناسایی الگوهای موجود در تراکنش ها و تعیین تراکنش های غیرعادی (مشکوک به تقلب) است. لیکن تشخیص به موقع تقلب ها و جلوگیری از وقوع آنها، به صورت دستی امکان پذیر نبوده و مستلزم صرف زمان و نیروی انسانی بسیاری خواهد بود. لذا، کشف دقیق و سریع تقلب های صورت پذیرفته در تراکنش های بانکی مبتنی بر کارتهای بانکی به صورت خودکار، امری ضروری به نظر می رسد.

این تحقیق درصدد شناسایی عوامل مؤثر بر تقلب و سوءاستفاده از کارتهای بانکی است تا بتوان در مورد برخی از عوامل مؤثر راهکارهایی را پیشنهاد نمود و با اطلاع رسانی و آموزش بهتر و مناسب تر به دارندگان این کارتها، تا حدودی از مسئله تقلب و سوءاستفاده جلوگیری نمود. امروزه محققین از تکنیک دسته بندی در داده کاوی به عنوان یکی از ابزارهای مورداستفاده در بحث شناسایی تقلب کارتهای اعتباری (بانکی) استفاده می کنند. لذا در این مقاله، در ابتدا به بررسی برخی از پژوهش های صورت گرفته در این حوزه پرداخته خواهد شد و در ادامه روش پیشنهادی معرفی و در بخش چهارم به بررسی نتایج حاصل از پیاده سازی روش پیشنهادی و در نهایت به مقایسه و نتیجه گیری کلی پرداخته خواهد شد.

۲-تحقیقات صورت گرفته در حوزه ی به کارگیری روشهای داده کاوی در تشخیص و پیش بینی کلاهبرداری بانکی

تقلب عبارت است از سوءاستفاده از سود یک سازمان بدون اینکه لزوماً به عواقب قانونی آن منجر شود [۴]. وانگ و هان (۲۰۱۸)، از الگوریتم ماشین بردار پشتیبان به همراه ترکیب با الگوریتم خوشه بندی k-means به منظور معرفی مدل پیش بینی تقلب کارت اعتباری براساس تجزیه و تحلیل خوشه ای و یکپارچه سازی استفاده کردند. در این پژوهش در ابتدا پیش پردازشی بر روی داده ها صورت گرفته و داده ها متوازن شده اند و بعد از آن در دو مرحله خوشه بندی و طبقه بندی عملیات پیش بینی تقلب صورت گرفته است، مقایسه بر اساس معیار دقت صورت گرفته است که در حالت ترکیب با خوشه بندی برابر با ۹۸٪ و در حالت استفاده از الگوریتم ماشین بردار به تنهایی برابر با ۹۷٪ است [۵].

سیو و همکاران (۲۰۱۷)، تحقیقی با عنوان کشف تراکنش های متقلب بانکی با استفاده از درخت تصمیم ارائه دادند. در این مقاله بیان شده است که خرید و بانکداری آنلاین با رشد اینترنت و با استفاده از کارت اعتباری افزایش یافته است. همراه با این توسعه، تعداد تقلب کارت اعتباری نیز افزایش یافته است. لیکن امروزه بسیاری از تکنیک های مدرن مبتنی بر هوش مصنوعی، در شناسایی معاملات جعلی مختلف کارت اعتباری تکامل یافته است. در این مقاله یک سیستم کشف تقلب در پردازش تراکنش کارت اعتباری با استفاده از یک درخت تصمیم گیری با ترکیبی از الگوریتم Luhn و الگوریتم هانت صورت گرفته است. الگوریتم Luhn برای اعتبارسنجی شماره کارت استفاده می شود. این الگوریتم تضمین نمی کند که آیا تراکنش متقلب یا واقعی است. اما اگر دو آدرس مطابقت داشته باشد، تراکنش ممکن است متقلب بوده و معامله مشکوک نامگذاری می شود. یک مشتری معمولا تراکنش های مشابهی را از نظر مقدار، انجام می دهد. از آنجا که تقلب کننده احتمالا با حساب کاربری متفاوت است، تراکنش او می تواند به عنوان استثناء شناسایی شود و باید بررسی گردد [۶].

بهرا و پانیگراهی (۲۰۱۷)، مقاله ای با عنوان کشف تراکنش های متقلب بانکی با استفاده از سیستم فازی ارائه دادند. در این مقاله یک

[4] Prepaid

سیستم فازی عصبی دو مرحله ای برای تشخیص تقلب کارت اعتباری پیشنهاد شده است. در این پژوهش یک تراکنش ورودی در ابتدا توسط یک سیستم تطبیق الگو در مرحله اول پردازش می شود. این جزء شامل یک فرآیند خوشه بندی فازی و یک فرآیند تطبیق آدرس است و هر یک از آنها بر مبنای میزان انحراف آن، یک معیار را به تراکنش اختصاص می دهد. یک سیستم استنتاج فازی یک میزان مشکوک را محاسبه می کند و تراکنش را به یکی از حالات واقعی ، مشکوک یا جعلی تقسیم می کند. هنگامی که یک تراکنش به عنوان مشکوک شناسایی شد، یک شبکه عصبی (که با انجام تراکنش های گذشته آموزش داده شده) به کار گرفته می شود تا تأیید کند که آیا این یک اقدام واقعی یا جعلی بوده است [۷].

فو و همکاران (۲۰۱۶)، از شبکه عصبی برای پیش بینی تقلب در استفاده از کارتهای اعتباری استفاده کرده اند. این پژوهش با سه معیار میزان حساسیت و معیار f و میزان دقت با الگوریتم های دیگر مقایسه شده است. این معیارها برای این پژوهش به ترتیب ۸۲ ٪ ، ۶۹ ٪ و ۷۹٪ است که به نسبت سه الگوریتم دیگر، بهبود پیدا کرده اند [۸].

محمدی والا و فرهودی نژاد (۱۳۹۶) در پژوهش خود به شناسایی تقلب در تراکنش های بانکی با استفاده از تکنیکهای داده کاوی پرداختند. در این تحقیق به تشخیص رفتارهای مشکوک مشتریان بانک بر اساس روشهای داده کاوی پرداخته شده است. مدل معرفی شده، ترکیبی از روش خوشه بندی و دسته بندی است. در ابتدا با استفاده از الگوریتم Two Step تراکنش های مشتریان برچسب گذاری شده و سپس با استفاده از آنها به دسته بندی تراکنش ها و تشخیص رفتارهای مشکوک با استفاده از شبکه عصبی RBF پرداخته شده است. نتایج تحقیق نشان می دهد که مدل پیشنهادی قادر است ۸۵٪ از تراکنش های مشکوک را به درستی تشخیص دهد که دقت بالایی در شناسایی تراکنش های مشکوک محسوب می شود.[۹]

حسن نژاد و توکلایی (۱۳۹۵) در پژوهش خود، به ارائه یک روش هوشمند برای شناسایی کلاهبرداری در کارتهای اعتباری پرداختند. در این تحقیق، با استفاده از روشهای مختلف داده‌کاوی به بررسی شناسایی کلاهبرداری در کارتهای اعتباری پرداخته شده است. روش پیشنهادی این تحقیق، شامل سه بخش عمده انتخاب مشخصه‌های مهم، تعیین استراتژی بهینه نمونه‌برداری و مدل‌سازی حساس به هزینه می‌باشد. در بخش نخست از روش پیشنهادی، از الگوریتم ژنتیک برای تعیین مشخصه‌های مهم استفاده شده است. در ادامه، با استفاده از روشی مبتنی بر طراحی آزمایشها، نسبت بهینه هر یک از دسته‌ها برای انجام نمونه‌برداری تعیین شده است. در بخش مدل‌سازی نیز از درخت تصمیم C4.5حساس به هزینه، به عنوان دسته‌بند پایه در الگوریتم آدابوست، استفاده شده است. در پایان، با استفاده از یک مجموعه داده واقعی نشان داده شده است که روش پیشنهادی تحقیق، با حداقل ۱۴٪ کاهش هزینه دسته‌بندی اشتباه، نتیجه بهتری نسبت به روشهای درخت تصمیم، بیزین ساده، شبکه بیزی، شبکه عصبی و سیستم ایمنی مصنوعی داشته است[۱۰].

خدابخشی و فرتاش(۱۳۹۵)، مقاله ای با عنوان یک روش مبتنی بر KNN جهت کشف تقلب در عملیات بانکداری، ارائه دادند. در این مقاله، از تکنیک K نزدیکترین همسایه و k-means جهت بهبود دقت الگوریتم کشف تقلب های صورت گرفته در تراکنش های مربوط به کارتهای اعتباری در سیستم بانکداری الکترونیک استفاده شده است. در نهایت نتایج بدست آمده از روش پیشنهادی از لحاظ دقت کشف تقلب های بانکی و سرعت شناسایی این تقلب ها با سایر روشها ، مورد مقایسه و ارزیابی قرار گرفته است[۱۱].

زارع‌پور و همکاران (۲۰۱۲)، در تحقیقی با عنوان شناسایی تقلب در کارت اعتباری بیان کرده اند که تقلب با کارت اعتباری، با پیشرفت تکنولوژی مدرن و تکنیک‌های ابر و با استفاده از ارتباطات جهانی، به‌طور قابل ملاحظه‌ای افزایش یافته است. کلاهبرداران به طور مداوم تلاش می‌کنند تا قوانین و تاکتیک‌های جدیدی را برای اعمال اقدامات غیرقانونی پیدا کنند. بدین ترتیب، سیستم‌های تشخیص تقلب برای بانکها و موسسات مالی ضروری است تا زیان آنها را به حداقل برساند. با این حال، به علت عدم وجود مجموعه داده‌های تراکنش‌های کارتهای اعتباری برای محققان، ادبیات منتشر شده در مورد تکنیک‌های تشخیص تقلب کارت اعتباری، وجود ندارد. در این پژوهش از الگوریتم‌های نایو بیز ، ماشین‌های بردار پشتیبانی و الگوریتم نزدیکترین همسایگی استفاده شده است. این تکنیک‌ها را می‌توان به تنهایی و یا با اشتراک با استفاده از تکنیک‌های دیگر برای ایجاد طبقه بندی‌ها مورد استفاده قرار داد. این پژوهش تنها از نظر زمان اجرایی با دیگر الگوریتم ها مقایسه شده است که در نهایت نتیجه بدست آمده نشان دهنده برتری زمان اجرا نسبت به روشهای دیگر است [۱۲].

سلطانی محمدی (۱۳۹۳) در پژوهش خود، به استفاده از روشهای داده‌کاوی در تشخیص و پیش‌بینی کلاه‌برداری‌های بانکی پرداخته است. او بیان کرده است که با استفاده از تکنیک‌های داده‌کاوی و با بررسی عمیق روی داده‌های خام در سطحی گسترده ، می‌توان به نتایج قابل تحلیل دست یافت. بررسی داده‌ها به روشهای مختلف و یافتن الگوی خاص و قابل تکرار در آنها، می‌تواند روش مناسبی برای کشف یا در درجه برتر آن، پیش‌بینی رفتار یا یک رخداد باشد [۱۳].

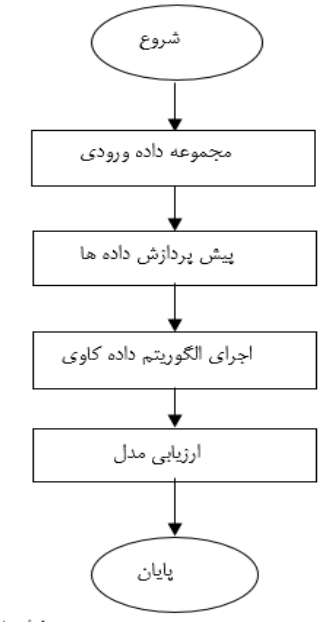
رهنمای رودپشتی (۱۳۹۱) به پژوهشی تحت عنوان داده کاوی و کشف تقلب های مالی پرداخت. تحقیق فوق، اثربخشی تکنیک های داده کاوی در تشخیص رفتارهای متقلبانه شرکتهایی که صورت های مالی متقلبانه گزارش نموده اند را بررسی کرده تا عوامل موثر بر اینگونه رفتارها را شناسایی کند. این پژوهش به روش شناخت تاریخی با بهره گیری از اسناد کتابخانه ای و به پشتوانه پیشینه و تحقیقات محققان، شواهدی جهت پاسخ به سوالات تحقیق ارائه می کند. نتایج مطالعه نشان می دهد که اولا، تکنیک های داده کاوی، در شناسایی صورتهای مالی متقلبانه سودمند هستند. ثانیا، داده کاوی، به عنوان کانون هدایت فکر در مدیریت کسب و کارها، جهت کشف تقلب می تواند مورد توجه قرار گیرد [۱۴].

۳-روش پیشنهادی

مجموعه داده مورد استفاده در این پژوهش، مربوط به

تراکنشهای مشتریان یک بانک تایوانی بوده که برخی از آنها در گروه مشتریان متقلب بانکی قرار گرفته اند [۱۵]. آخرین ویژگی در هر رکورد در این مجموعه داده، ویژگی کلاس است که نمایش دهنده افراد متقلب یا افراد عادی است. ویژگیهای موجود در هر رکورد این مجموعه داده شامل موارد زیر هستند: ۱-میزان اعتبارات دریافتی ، ۲-هدف از دریافت تسهیلات توسط مشتری ،۳-جنسیت مشتری ۴- وضعیت تاهل مشتری ، ۵- میزان تحصیلات مشتری ، ۶-سن مشتری ۷-وضعیت پرداختهای ماهیانه یکسال قبل مشتری از بابت تاخیر در پرداخت ، ۸-مبلغ اقساط ماهیانه ، ۹-مبلغ پرداختی متغیر ماههای قبل ، ۱۰-نوع تسهیلات دریافتی .(مانند تسهیلات بیمه عمر و...) ، ۱۱-درصد نرخ قسط در درآمد قابل تصرف ، ۱۲-شغل مشتری ، ۱۳-کلاس مشتری.

در این پژوهش از ترکیب الگوریتم بگینگ و خوشه بندی kmeans استفاده شده است. با توجه به این مسئله که آدابوست یک الگوریتم چندگانه است بنابراین با الگوریتم های دیگر توانایی ترکیب را خواهد داشت که الگوریتم های پایه ترکیب شده نیز در ادامه معرفی می شوند. شکل ۱ فلوچارت روند پیشنهادی را نمایش می دهد.



شکل۱: فلوچارت روند پیشنهادی

۳-۱- ورود داده ها و پیش پردازش داده ها

داده های خام ورودی در ابتدا نیاز به پیش پردازش دارند، بنابراین در مرحله اول برخی از عملیات پیش پردازشی بر روی آنها صورت خواهد گرفت. در این مجموعه داده، برخی از فیلدها فاقد مقدار هستند که این رکوردها ارزش نگهداری را نخواهند داشت، بنابراین این داده ها در مجموعه داده حذف خواهند شد.

۳-۲- اعمال الگوریتم خوشه بندی بر روی داده ها

در مرحله دوم، از الگوریتم خوشه بندی به منظور گروه بندی داده ها در خوشه های مناسب استفاده شده است. با استفاده از خوشه بندی بر روی این مجموعه داده ها می توان افرادی که متقلب و یا فرد عادی بوده اند اما ویژگیهای متفاوتی داشته اند را در خوشه های یکسان قرار داد و با کنار هم قرار دادن ویژگیهای متفاوت افراد، طراحی مدل را ارتقا بخشیده و در نتیجه دقت شناسایی سیستم را افزایش داد.

۳-۳- تقسیمات مجموعه داده

مجموعه داده ها به دو قسمت آموزشی و تست تقسیم می شوند. این تقسیمات به ۳ گروه ۷۰ ٪ آموزشی و ۳۰ ٪ تست، ۸۰ ٪ آموزشی و ۲۰ ٪ تست و ۹۰ ٪ آموزشی و ۱۰ ٪ تست بوده اند.

۳-۴- الگوریتم ترکیبی

در این مرحله از پژوهش، نوع الگوریتم بگینگ و الگوریتم های ترکیب شده با آنها انتخاب خواهند شد. الگوریتم های ترکیب شده شامل الگوریتم های J48 ، Logestic و Random Forest می باشند. با توجه به این مسئله که الگوریتم های انتخاب شده، الگوریتم های ترکیبی مانند آدابوست و بگینگ هستند (الگوریتم های ترکیبی برای بالا بردن دقت الگوریتم های پایه استفاده می شوند به عنوان مثال، الگوریتم آدابوست، طبقه بند در هر مرحله به نفع نمونه های غلط طبقه بندی شده در مراحل قبل، تنظیم می گردد، بنابراین با این روند رویکرد الگوریتم پایه را ارتقا خواهد داد) و کاربرد آنها در ارتقا الگوریتم های پایه است و با توجه به نوع داده ها که به صورت عددی هستند و این که درختها بر روی داده های عددی دقت بالاتری دارند ، بنابراین از الگوریتم J48 استفاده شده است.

۳-۵- ارزیابی مدل

در این قسمت از روند پیشنهادی، مدل با استفاده از معیارهای ارزیابی دقت، حساسیت، درصد پاسخ بر روی نمونه های منفی و با استفاده از نمونه های آزمایشی تست شده و نتایج آن با دیگر الگوریتم ها مقایسه خواهد شد.

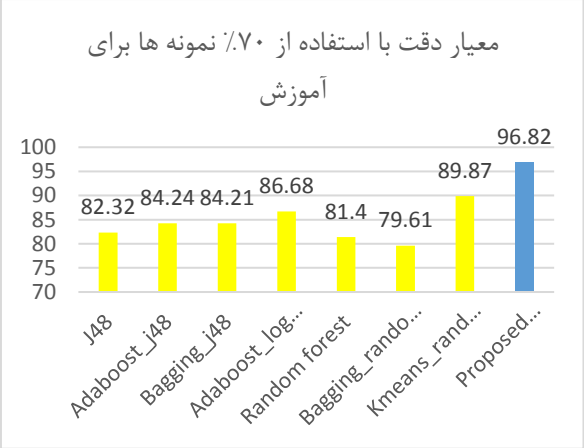
۴-بحث و نتایج

مجموعه داده مورد استفاده، داده های کلاسه بندی شده از مجموعه داده یکی از بانکهای تایوان است که فیلدهای مشخصی دارد [۱۵]. الگوریتم های استفاده شده بر روی این مجموعه داده، الگوریتم ترکیبی با الگوریتم های پایه هستند که نتایج انواع ترکیب و اجرای تک تک الگوریتم های پایه، معرفی می گردند.

نتایج الگوریتم های استفاده شده، به صورت جداگانه با مقادیر آموزشی ۷۰، ۸۰ و ۹۰ درصد نمایش داده شده است. این الگوریتم ها به

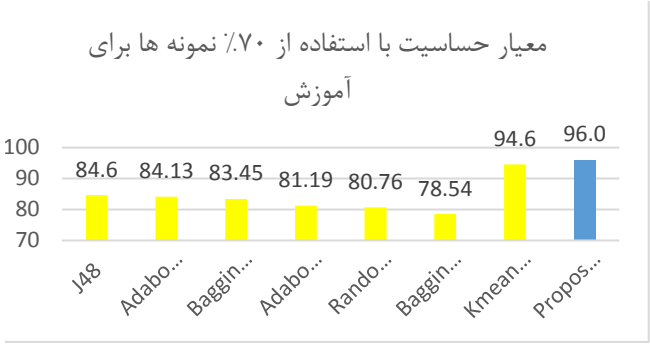
صورت پایه و ترکیبی استفاده شده اند و نتایج آنها در قالب نمودار برای معیارهای متفاوت نمایش داده شده است. در شکل۲، نتایج معیار دقت (Accuracy) روش پیشنهادی بر روی مجموعه داده، با تعدادی از الگوریتم های طبقه بندی دیگر مانند Random forest ، J48 و با مجموعه داده آموزشی ۷۰ ٪ نمایش داده شده است. میزان دقت در الگوریتم هایی که از روند ترکیبی آدابوست استفاده کرده اند به طور میانگین، برابر با ۸۵ ٪ است.

برای ایجاد یک مدل مناسب در روند طبقه بندی داده ها، استفاده از خوشه بندی می تواند در ایجاد مدل، توانایی تشخیص نمونه ها را بالا ببرد. یعنی در ابتدا با استفاده از خوشه بندی، داده های مشابه گروه بندی می شوند. بنابراین در مرحله ایجاد مدل، قوانین ایجاد شده خوشه های مجزاتری را ایجاد خواهند کرد، بنابراین میزان تشخیص برچسب متقلب بودن و یا فرد عادی بودن افزایش پیدا خواهد کرد که میزان دقت در آن برابر با ۹۶٪ شده است.



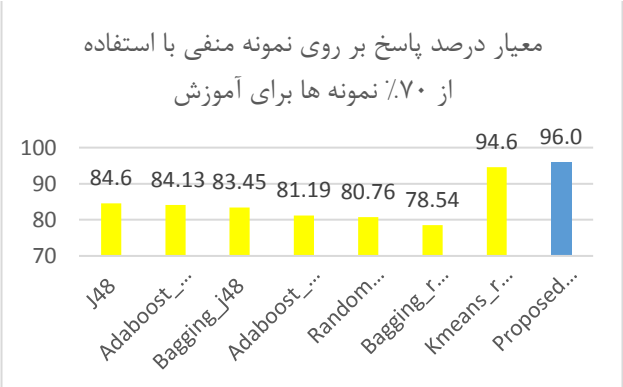
شکل ۲ : نتایج معیار دقت روند پیشنهادی و سایر الگوریتم های داده کاوی با مجموعه داده ۷۰ ٪ آموزشی

در پژوهش های داده کاوی، در نظر گرفتن معیار دقت به تنهایی کافی نیست و باید از معیارهای دیگر نیز برای ارزیابی استفاده شود. در شکل۳ مقایسه نتایج معیارحساسیت روند پیشنهادی و الگوریتم های طبقه بندی دیگر با مجموعه داده ۷۰ ٪ نمایش داده شده است. معیار حساسیت ،درصد تشخیص نمونه های مثبت را نشان داده و در واقع نمایش دهنده عملیاتی است که تقلبی در آن اتفاق نیفتاده است. میزان معیار حساسیت برای روند پیشنهادی برابر با ۹۶ ٪ است. همان طور که نتایج دیگر الگوریتم ها نشان داده است، الگوریتم پیشنهادی نتایج مناسب تری را نسبت به دیگر الگوریتم ها تولید کرده است.



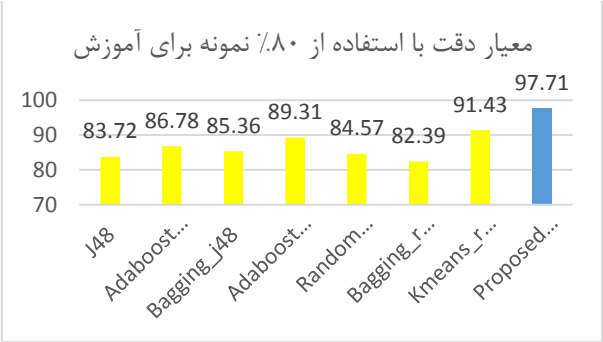
شکل ۳ : نتایج معیار حساسیت روند پیشنهادی و سایر الگوریتم های داده کاوی با مجموعه داده ۷۰٪ آموزشی

در شکل ۴ مقایسه نتایج معیار درصد پاسخ بر روی نمونه منفی (Specificity) روند پیشنهادی و الگوریتم های داده کاوی دیگر نمایش داده شده است، این معیار برای روند پیشنهادی برابر با ۹۸ ٪ است. این معیار نشان دهنده این است که این الگوریتم توانسته است درصد پاسخ صحیح بالایی بر روی نمونه های منفی، یعنی عملیات متقلبانه، داشته و نتایج بهتری نسبت به دیگر الگوریتم ها تولید کند. در الگوریتم های ترکیبی مانند رویکرد الگوریتم استفاده شده، مدلهای ایجاد شده خاص داده های آموزشی نیست و بنابراین نتایج مناسبی در پیش بینی صورت خواهد گرفت. دلیل خاص نبودن این مدل ها، به علت نمونه گیری های متفاوت صورت گرفته در این داده هاست که با نمونه گیری های متفاوت و ایجاد مدل با آنها تمامی جوانب مجموعه داده ها در نظر گرفته شده است. بنابراین میزان معیارهای بدست آمده نیز بالاتر خواهد بود.

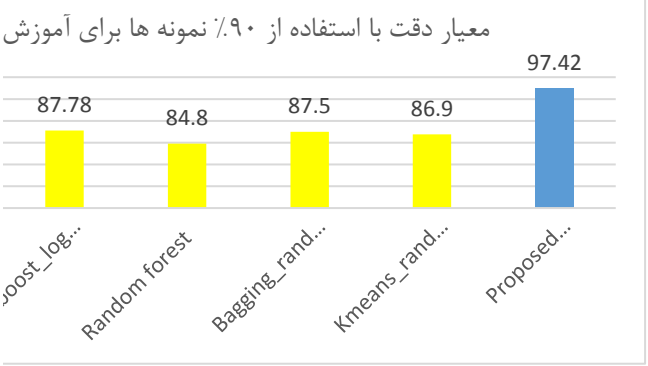


شکل ۴ : نتایج معیار درصد پاسخ بر روی نمونه منفی روند پیشنهادی و سایر الگوریتم های داده کاوی با مجموعه داده ۷۰ ٪ آموزشی

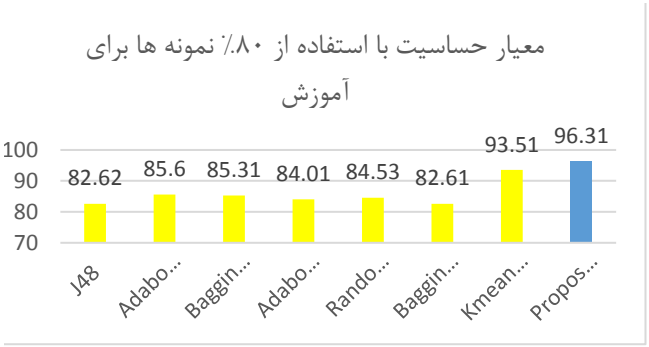
تمامی عملیات انجام گرفته در بخش قبل، مجددا و این بار با ۸۰ ٪ نمونه برای آموزش و ۲۰٪ نمونه برای آزمایش انجام و نتایج حاصل در شکلهای ۵ الی ۷ نشان داده شده است. نتایج حاصل مجددا برتری روش پیشنهادی را نسبت به سایر روشهای طبقه بندی نشان می دهد.



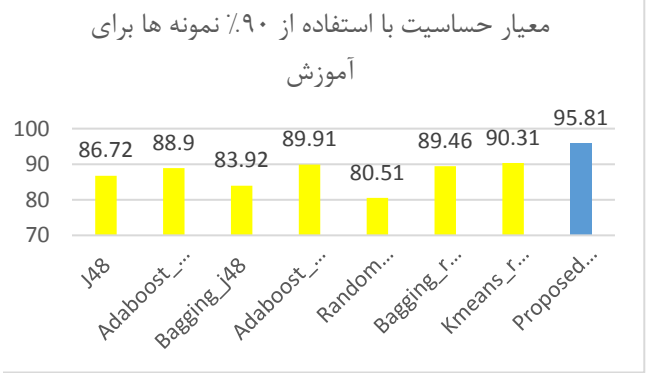
شکل ۵: نتایج معیار دقت روند پیشنهادی و سایر الگوریتم های داده کاوی با مجموعه داده ۸۰ ٪ آموزشی



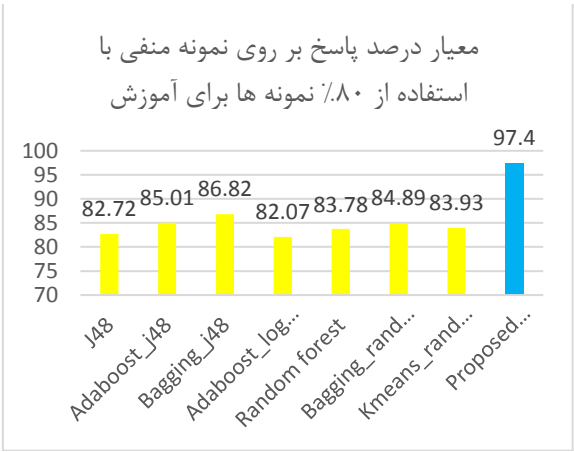
شکل ۸ : نتایج معیار دقت روند پیشنهادی و سایر الگوریتم های داده کاوی با مجموعه داده ۹۰ ٪ آموزشی



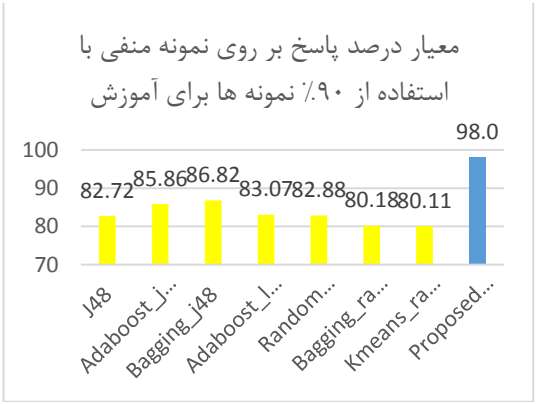
شکل ۶ : نتایج معیار حساسیت روند پیشنهادی و سایر الگوریتم های داده کاوی دیگر با مجموعه داده ۸۰٪ آموزشی



شکل ۹ : نتایج معیار حساسیت روند پیشنهادی و سایر الگوریتم های داده کاوی با مجموعه داده ۹۰ ٪ آموزشی



شکل ۷ : نتایج معیار درصد پاسخ بر روی نمونه منفی روند پیشنهادی و سایر الگوریتم های داده کاوی با مجموعه داده ۸۰ ٪ آموزشی

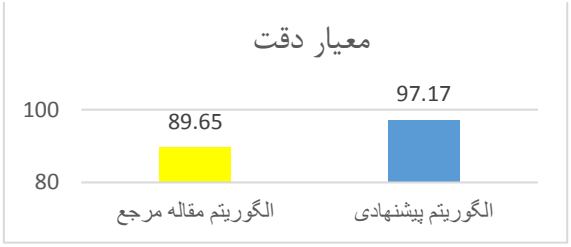


شکل ۱۰ : نتایج معیار درصد پاسخ بر روی نمونه منفی روند پیشنهادی و سایر الگوریتم های داده کاوی با مجموعه داده ۹۰ ٪ آموزشی

مجددا تمامی عملیات انجام گرفته در بخش قبل، و این بار با ۹۰ ٪ نمونه برای آموزش و ۱۰٪ نمونه برای آزمایش انجام و نتایج حاصل در شکلهای ۸ الی ۱۰ نشان داده شده است. نتایج حاصل مجددا برتری روش پیشنهادی را نسبت به سایر روشهای طبقه بندی نشان می دهد.

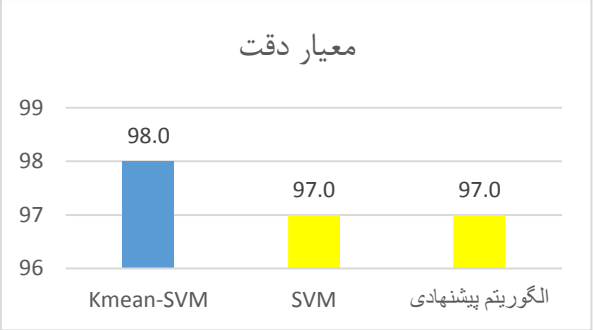
به منظور ارزیابی عملکرد روش پیشنهادی، لذا نتایج این پژوهش، با دو مقاله دیگر در همین حوزه مقایسه شده است. نتیجه الگوریتم پیشنهادی در معیار دقت، با الگوریتم مقاله وج و همکاران

(۲۰۱۷) که از یک الگوریتم ماشین بردار پشتیبان استفاده کرده است، در شکل ۱۱ مقایسه شده است. همانطور که دیده می شود، نتایج الگوریتم پیشنهادی به میزان ۷/۵۲ ٪ از تحقیق فوق بهتر بوده که این نتیجه بهتر، به دلیل انجام خوشه بندی دقیقتر اولیه است.

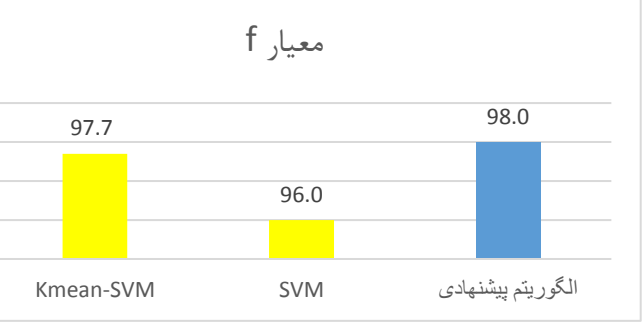


شکل ۱۱ : مقایسه نتایج معیار دقت در الگوریتم پیشنهادی و تحقیق وج و همکاران [۱۶]

مقایسه نتایج الگوریتم پیشنهادی با الگوریتم مقاله وانگ و هان (۲۰۱۸)، در شکلهای ۱۲ و ۱۳ نمایش داده شده است. در این مقاله از ترکیب الگوریتم خوشه بندی kmeans و svm استفاده شده است. نتایج بدست آمده از الگوریتم ماشین بردار پشتیبان به اندازه ۱٪ از نتایج الگوریتم پیشنهادی بالاتر است و دلیل این برتری می تواند به علت عملیات پیش پردازشی انتخاب ویژگی صورت گرفته بر روی مجموعه داده توسط آن محققین باشد.



شکل ۱۲ : مقایسه نتایج معیار دقت در الگوریتم پیشنهادی و تحقیق وانگ و هان [۵]



شکل ۱۳ : مقایسه نتایج معیار f در الگوریتم پیشنهادی و تحقیق وانگ و هان [۵]

مقایسه نتایج الگوریتم پیشنهادی و دو مقاله مرجع وانگ و هان [۵] و وج و همکاران [۱۶] نشان می دهد که در برخی موارد، الگوریتم پیشنهادی به دلیل انتخاب خوشه بندی، نتایج مناسب تری داشته و در نتیجه معیار دقت بالاتری بدست آمده است.

۵- نتیجه‌گیری

در این پژوهش به بررسی کشف تقلب در استفاده از کارتهای بانکی در تراکنشهای مالی با استفاده از تکنیک های داده کاوی و با بهره گیری از داده های یک بانک تایوانی ، پرداخته شده است. رویکرد پیشنهادی، استفاده از الگوریتم های ترکیبی بگینگ است که با الگوریتم های پایه ترکیب شده است. در رویکرد پیشنهادی از خوشه بندی نیز استفاده شده است. به منظور بررسی ترکیبهای مختلف الگوریتم، الگوریتم های پایه J48 و Random forest استفاده شده و در کنار آنها، تمامی ترکیبهای متفاوت از این گروه ها صورت گرفته است و نتایج آنها در قالب نمودار ارائه شده است.

یکی از عوامل موفقیت ترکیب پیشنهادی نسبت به رویکردهای دیگر، استفاده از الگوریتم خوشه بندی است. به دلیل نزدیک بودن مشخصات افراد متقلب و غیر متقلب ، مجموعه داده ها نیاز به خوشه بندی و گروه بندی دارد. دلیل دیگر بهبود در این رویکرد، استفاده از الگوریتم bagging است که بر روی نمونه های منفی، به طور مناسبتری عمل می کند. این الگوریتم به علت تقویت مدل برای نمونه های منفی، جوابهای مناسبی را ایجاد کرده است.

۶- مراجع

1. Jain, N., & Khan, V. (2018). Credit Card Fraud Detection using Recurrent Attributes. People, 5(2).

۲. زارع دستجردی،آذر .(۱۳۹۲). کشف و استخراج الگوهای ناهنجاری در کارت های پرداخت الکترونیک، پایان نامه کارشناسی ارشد رشته هوش مصنوعی، دانشکده تربیت معلم .

3. Awoyemi, J., Adetunmbi, A., & Oluwadare, S. (2018). Effect of Feature Ranking on Credit Card Fraud Detection. In *2nd International Conference on Information and Communication Technology and Its Applications (ICTA 2018)*, pp. 140-147.

4. Montazer, G. A., & ArabYarmohammadi, S. (2015). Detection of phishing attacks in Iranian e-banking using a fuzzy–rough hybrid system. *Applied Soft Computing*, 35, 482-492.

5. Wang, C., & Han, D. (2018). Credit card fraud forecasting model based on clustering analysis and integrated support vector machine. *Cluster Computing*, 1-6.

6. Save, P., Tiwarekar, P., Jain, K. N., & Mahyavanshi, N. (2017). A novel idea for credit card fraud detection

using decision tree. *International Journal of Computer Applications*, *161*(13).

7. Behera, T. K., & Panigrahi, S. (2017). Credit Card Fraud Detection Using a Neuro-Fuzzy Expert System. *In Computational Intelligence in Data Mining* (pp. 835-843). Springer, Singapore.

8. Fu, K., Cheng, D., Tu, Y., & Zhang, L. (2016, October). Credit Card Fraud Detection Using Convolutional Neural Networks. In *International Conference on Neural Information Processing* (pp. 483-490). Springer International Publishing.

۹. محمدی والا, حمید، فرهودی نژاد،اکبر. (۱۳۹۶)، شناسایی تقلب در تراکنش های بانکی با استفاده از داده کاوی: مطالعه موردی شناسایی در تراکنش های بانک مهراقتصاد، مهندسی کامپیوتر و پژوهشهای نیاز محور آخرین دستاوردهای در فناوری اطلاعات، مشهد، موسسه آموزش عالی خاوران.

۱۰. حسن نژاد، سینا، توکلایی، حمید. (۱۳۹۵). ارائه روش های داده کاوی جهت تشخیص کلاهبرداری در کارت های اعتباری، دومین همایش ملی مهندسی رایانه و مدیریت فناوری اطلاعات، تهران، گروه پژوهشی بوعلی.

۱۱. خدابخشی, مسعود، فرتاش، مهدی، (۱۳۹۵)، یک روش مبتنی بر KNN جهت کشف تقلب در عملیات بانکداری، *سومین کنگره بین المللی کامپیوتر، برق و مخابرات*، تربت حیدریه، دانشگاه تربیت حیدریه.

12. Zareapoor, M., Seeja, K. R., & Alam, M. A. (2012). Analysis on Credit Card Fraud Detection Techniques: Based on Certain Design Criteria. *International Journal of Computer Applications*, 52(3).

۱۳.سلطانی محمدی،آرمان. (۱۳۹۲). بکار گیری روش‌های داده‌کاوی در تشخیص و پیش‌بینی کلاه‌برداری‌های بانکی در ایران، پایان‌نامه کارشناسی ارشد رشته مهندسی فن‌آوری اطلاعات، دانشگاه قم.

۱۴. رهنمای رودپشتی، فریدون. (۱۳۹۱). داده کاوی و کشف تقلب های مالی. دانش حسابداری. ۱(۳): ۱۷-۳۳.

15. Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, *36*(2), 2473-2480.

16. Wedge, R., Kanter, J. M., Rubio, S. M., Perez, S. I., & Veeramachaneni, K. (2017). Solving the" false positives" problem in fraud prediction. *arXiv preprint arXiv:1710.07709*.