کشف تقلبات در کارت های اعتباری با استفاده از شبکه عصبی و درخت تصمیم

رضا روشنی
۱. موسسه آموزش عالی غیر انتفاعی لامعی گرگانی
۲. گروه مهندسی کامپیوتر، دانشگاه ملی مهارت، تهران، ایران

معصومه کیائی
۱. موسسه آموزش عالی غیر انتفاعی لامعی گرگانی

چکیده
استفاده از کارت‌های اعتباری برای جلوگیری از حوادث ناشی از بهمراه داشتن پول نقد رواج پیدا کرده است. اما این موضوع نه تنها باعث عقب نشینی افراد سودجو نشده، بلکه راه‌های جدید با خطر کمتری را پیش روی آن‌ها گشوده است. از بررسی آمارهای بانکی و سایر متولیان کارت‌های اعتباری چنین بر می‌آید که تقلب در کارت‌های اعتباری بطور فزاینده‌ای در حال افزایش است و با توجه به حجم بالای تراکنش‌های صورت گرفته، نیاز به ابزار مناسبی برای کشف تقلب در داده‌ها احساس می‌شود. داده‌کاوی علم جدیدی است که در این زمینه مطرح شده، بطوریکه با استفاده از تکنیک‌ها و الگوریتم‌های آن می‌توان در این حجم انبوه داده به پردازش پرداخته و موارد مشکوک را جستجو و شناسایی نمود. در این پژوهش به منظور ارائه روشی در زمینه کشف تقلبات صورت گرفته در تراکنش‌های کارت‌های اعتباری به بررسی دو تکنیک پر کاربرد داده‌کاوی یعنی درخت تصمیم و شبکه عصبی بر روی دو مجموعه داده پرداخته می‌شود و با استفاده از ابزار داده‌کاوی متلب این الگوریتم‌ها در دو مرحله بر روی مجموعه داده‌های انتخابی پیاده سازی خواهند شد و نتایج آن‌ها مورد بررسی و مقایسه قرار خواهد گرفت. نتایج پیاده‌سازی نشان می‌دهد که الگوریتم درخت تصمیم نسبت به شبکه عصبی درصد بیشتری از معیارها را به خود اختصاص داده است.

واژگان کلیدی: کارت اعتباری، کشف تقلب، داده‌کاوی، درخت تصمیم، شبکه عصبی.

۱. مقدمه

یکی از بزرگترین سازمان‌هایی که با مشتریان به صورت کاملا مستقیم در تعامل هستند بانک‌ها می‌باشند و با توسعه روز افزون تکنولوژی، و جایگزینی سیستم‌های نوین در تمامی عرصه‌ها ، بجای ساختار سنتی و کهنه، نظام بانکداری نیز پیرو این ارتقاء، سعی برآن دارد که در حوزه فن آوری، مطلوبیت ممکن را برای مشتریان فراهم آورد. یکی از ابزارهای مهم در این عرصه، کارت اعتباری[1] است که حجم بالای نقل و انتقالات مالی افراد حقیقی و حقوقی را شامل شده است. کارت‌های اعتباری می‌توانند زندگی را آسان‌تر کنند. آن‌ها به مشتریان این امکان را می‌دهند که از اعتبار خود در زمان، مکان و میزان دلخواه، بدون حمل پول نقد استفاده کنند. چالش‌های امنیتی مختلف به دلیل محبوبیت استفاده از کارت‌های اعتباری در حال افزایش است و همین موضوع باعث تشدید کلاهبرداری با هدف به دست آوردن منافع مالی غیر مجاز شده است. تقلب کارت اعتباری به وضعیتی اطلاق می‌شود که متقلب برای رفع نیازهای خود از کارت اعتباری دیگری استفاده می‌کند در حالی که صاحب آن کارت اعتباری از این امر آگاه نیست. روش‌های مختلفی برای کشف تقلب در کارت‌های اعتباری معرفی شده‌اند. با این حال، کلاهبرداران نیز دائما روش‌های جدیدی برای سرقت اطلاعات ارائه می‌دهند [۱۱]. در حال حاضر اجرای یک سیستم کشف تقلب کارآمد چالش اصلی این موضوع است. روش‌های جدیدی برای کشف تقلب وجود دارد؛ اما از داده‌کاوی[2] که هدف آن استخراج اطلاعات مورد علاقه از مجموعه‌ی گسترده‌ای از داده‌هاست به طور گسترده‌ای به عنوان ابزار تصمیم گیری فعال، استفاده می‌شود.

۱/۱. کارت اعتباری

در فرهنگ آکسفورد، کارت اعتباری چنین تعریف شده است:" کارتی که بانک یا غیر آن در اختیار مشتری قرار می‌دهد و او می‌تواند کالا و خدمات مورد نیاز خود را دریافت کند. "
بنابراین کارت اعتباری کارتی است که بانک‌ها و موسسات مالی و اعتباری خصوصی در اختیار مشتریان خود قرار می‌دهند تا او بتواند بهای کالا و خدمات مورد نیاز خود را بدون دغدغه پول فیزیکی با استفاده از دستگاه‌های مربوطه نظیر پایانه فروش (POS)[3]، دستگاه خودپرداز(ATM)[4] و یا اینترنت پرداخت کند[۱۲].

۱/۲. تقلب کارت‌های اعتباری

معمولا، تقلب به استفاده غیرمجاز و غیر قانونی از تسهیلات اعتباری یک حساب قانونی اشاره دارد. تقلب می‌تواند به عنوان یک فعالیت بزهکارانه تعریف شود که با نمایش‌های غیرقانونی برای بدست آوردن یک امتیاز ناعادلانه سروکار دارد[۱۳]. امروزه دزدی با کارت اعتباری بسیار امن‌تر، راحت‌تر و پر سودتر از دزدی با اسلحه است. صدمات به تجارت‌ها و بانک‌ها از طریق تقلب کارت‌های اعتباری با افزایش روز افزونی رو به روست.
بیان کردن وسعت تقلب کارت اعتباری مشکل است، چون ارقام در طول زمان تغییر می‌کنند (محتملا رشد می‌کنند). تا حدی که اغلب اوقات شرکت‌ها از منتشر ساختن ارقام تقلب بی میل هستند، به دلیل اینکه آن‌ها افکار عمومی را به وحشت می‌اندازند.

۱/۳. انواع تقلب در کارت‌های بانکی و اعتباری[۱]و[۲]

تقلب در کارت‌های اعتباری به طرق مختلفی امکان‌پذیر می‌باشد که در ذیل به آن‌ها اشاره شده است.

[1] Credit Card
[2] Data Mining
[3] Point of Sale
[4] Automated teller machine

- تقلب ورشکستگی

در این روش فرد با آگاهی از اینکه توان پرداخت بدهی خود را ندارد باز هم با کارت اعتباری خود معاملات انجام می‌دهد. این بدهی‌ها برای بانک از طرف فرد خاطی غیرقابل وصول است و خود بانک متحمل ضرر می‌شود. تنها راه جلوگیری از این تقلب‌ها بررسی وضعیت اعتباری دارنده حساب‌ها از طریق موسسات سنجش اعتباری است.

- کارت گمشده یا دزدی

در این روش کارت فردی مفقود یا دزدیده شده است و تا زمان بلوکه شدن حساب، از کارت برداشت‌هایی صورت می‌گیرد.

- تقلب درخواست

در این نوع تقلب، شخص به سه روش با دادن اطلاعات نادرست، کارت اعتباری دریافت می‌کند:

۱. هویت جعلی : شخصی متقلب با بدست آوردن اطلاعات غیرقانونی از شخص دیگر به افتتاح حساب کاربری با نام آن‌ها دست می‌زند.
۲. تقلب مالی : شخص به هنگام افتتاح حساب درباره وضعیت مالی خود اطلاعات نادرست ارائه می‌دهد.
۳. نرسیدن بسته پستی : بسته پستی حاوی کارت ارسال شده برای فرد در بین راه دزدیده می‌شود.

- کنترل حساب

شخص متقلب در این روش اطلاعات حساب مشتری خاصی را بدست آورده و می‌تواند حساب او را کنترل نماید.

- جعل کارت

در این روش مجرمان با روش‌های مختلفی مثل پاک کردن نوارهای مغناطیسی کارت بوسیله میدان مغناطیسی قوی یا دستکاری اطلاعات و مشخصات کارت و تطبیق آن با اطلاعات یک کارت قانونی یا کپی کردن داده‌های اصلی نوار مغناطیسی یک کارت بصورت الکترونیکی بر روی کارت دیگر یا حتی استفاده از کارت‌های سفید که نیاز به تایید برخی پایانه‌ها ندارد، به جعل کارت می‌پردازند.

- تبانی فروشنده

در این روش فروشنده خاطی اطلاعات مشتریان خود را در اختیار مجرمان قرار می‌دهد.

- تقلب اینترنتی

این نوع تقلب انواع مختلفی دارد:

۱. وب سایت‌های مشابه : شبیه سازی صفحه اینترنتی با فرم خرید از یک فروشگاه معتبر. پس از انجام خرید، اطلاعات کارت خریدار در اختیار مجرم است.
۲. وب سایت تقلبی فروش: سایت‌هایی راه اندازی شده که در آن کالاهای تقلبی با درصد تخفیف‌های بالا به حراج گذاشته می‌شود که ظاهر یک سایت حراج قانونی را دارد.

مولد کارت اعتباری: برنامه‌های کامپیوتری هستند که می‌توانند شماره انواع کارت‌های اعتباری مانند ویزا و مسترکارت را تولید کنند.

۱/۴. معرفی داده‌کاوی

داده‌کاوی عبارت است از کشف ارتباطات یا الگوهای معنی دار در یک پایگاه داده بسیار بزرگ و پیچیده. به عبارتی دیگر داده‌کاوی یک فرآیند خودکار یا نیمه خودکار استخرج دانش در قالب الگوهای پنهان از مجموعه اطلاعات ورودی می‌باشد. در داده‌کاوی ورودی عمدتا بسیار حجیم و پردازش دستی آن ناممکن است. در نتیجه نتایج حاصل از داده‌کاوی، با روش‌های سنتی پردازش اطلاعات، قابل دستیابی نمی‌باشد[۳].

اکتشاف دانش در پایگاه‌های داده یک فرآیند پیچیده است که شامل سه مرحله کلی آماده سازی داده‌ها، یادگیری مدل و ارزیابی و تفسیر مدل است. این سه مرحله در تمام پروژه‌های داده‌کاوی وجود دارند و اساس داده‌کاوی بر این سه مرحله استوار است.

آماده سازی داده اولین و مهم‌ترین مرحله در فرآیند داده‌کاوی است و هدف آن تأمین ورودی مناسب برای مرحله حیاتی یادگیری مدل است. به طور کلی به مجموعه عملیاتی که منجر به تولید مجموعه‌ای از داده‌های پالایش شده قابل کاوش خواهد شد اصطلاحاً آماده سازی داده می‌گویند و این عملیات عبارت اند از استخراج داده و پیش پردازش داده [۴].

در مرحله یادگیری مدل، نظم حاکم بر داده‌های پیش‌پردازش شده، با توجه به روش کاوش داده‌ای که انتخاب می‌شود، شناسایی شده و مدل تولید شده برای ارزیابی به مرحله بعد یعنی ارزیابی و تفسیر مدل منتقل خواهد شد. میتوان روش‌های مختلف کاوش داده را در دو گروه روش‌های پیش بینی(یادگیری باناظر) و روش‌های توصیفی(یادگیری بدون ناظر) طبقه بندی نمود. انواع الگوریتم‌ها و روش‌ها برای یادگیری مدل وجود دارند که در این مقاله ما بر روی الگوریتم شبکه عصبی و درخت تصمیم تمرکز خواهیم کرد و در ادامه توضیحات تکمیلی این دو الگوریتم ارائه می‌شود.

منظور از ارزیابی دانش آن است که می‌بایست میزان صحت دانش تولیدشده مشخص شود تا بتوان به آن اعتماد نمود و به صورت عملی از آن استفاده کرد. تفسیر مدل به معنای آن است که دانش تولید شده را مورد بررسی قرار داده و توجیهی معنایی برای تبیین منطق آن ارائه نماییم[۴].

۱/۴/۱. شبکه‌های عصبی

شبکه عصبی مصنوعی، از مجموعه نورون‌های عصبی تشکیل می‌شود که بصورت موازی با هم برای حل یک مسئله مشخص کار می‌کنند. شبکه‌های عصبی با مثال کار می‌کنند و نمی توان آن‌ها را برای انجام وظیفه‌ی خاص برنامه‌ریزی کرد. مثال‌ها باید با دقت انتخاب شوند در غیر این صورت زمان سودمند، تلف می‌شود و یا حتی بدتر از آن شبکه ممکن است نادرست کار کند. امتیاز شبکه عصبی این است که خودش کشف می‌کند که چگونه مسئله را حل کند، عملکرد آن غیرقابل پیشگویی است. مهمترین عواملی که گونه‌ها و کاربردهای شبکه عصبی را از یکدیگر متمایز می‌سازند، نوع نورون به کار گرفته شده، چیدمان یا معماری شبکه، بازه ورودی/خروجی هاست. وضعیت نسبی سلول‌ها در شبکه (تعداد و گروه بندی و نوع اتصالات آن‌ها) را معماری شبکه گویند. در معماری یک شبکه تعداد لایه‌ها و اتصالات بین آن‌ها مهم است. ورودی‌های شبکه با نام "لایه ورودی" و خروجی شبکه با نام "لایه خروجی" و در صورت نیاز، لایه های میان این دو " لایه پنهان " نامیده می‌شود[۱۴].

- لایه ورودی

این لایه ورودی‌ها را دریافت می‌کند و بر حسب قدرت ارتباطش با لایه بعد سیگنال ورودی را به لایه بعد می‌فرستد. قدرت ارتباط هر نورون با نورون دیگر را وزن آن نورون می‌گویند.

- لایه میانی

تعداد لایه‌های میانی و تعداد نورون‌های آن دلخواه است. لایه‌های میانی باید با دقت انتخاب شوند تا خروجی مناسب را به ما بدهند.

- لایه خروجی

گروه دیگری از نورون‌ها نیز از طریق خروجی‌های خود، جهان خارج را می‌سازند.

شبکه عصبی مصنوعی مانند یک تابع عمل می‌کند که به تعداد نورون‌های ورودی، ورودی می‌گیرد و به تعداد نورون‌های خروجی، خروجی می‌دهد.

- یادگیری در شبکه عصبی

هر شبکه عصبی سه مرحله آموزش، اعتبار سنجی و اجرا را پشت سر می‌گذارد. آموزش دیدن شبکه‌های عصبی در واقع چیزی جز تنظیم وزن‌های ارتباطی این نورون‌ها به ازای دریافت مثال های مختلف نیست تا خروجی شبکه به سمت خروجی مطلوب همگرا شود.

یک شبکه عصبی با تعیین رابطه بین ورودی‌ها و خروجی‌ها یاد می‌گیرد. این رابطه با محاسبه اهمیت نسبی ورودی‌ها و خروجی‌های سیستم تعیین می‌شود. با توجه به آزمون و خطا، سیستم نتایجش را با نتایج داده شده توسط متخصص مقایسه می‌کند، تا زمانی که به سطح مشخصی از دقت برسد. با هر آزمایش وزن‌های نسبت داده شده به ورودی‌ها تغییر می‌کند تا نتایج مطلوب به دست آید.

- مراحل یادگیری

۱-مقداردهی اولیه: وزن‌های اولیه به تصادف در محدوده مورد نظر انتخاب می‌شوند.

۲-فعال سازی: نورون‌ها با به کار بردن ورودی‌ها و خروجی‌های مطلوب فعال می‌شوند. خروجی واقعی با تکرار محاسبه می‌شود.

۳-آموزش وزن: وزن‌ها با توجه به خروجی مطلوب تصحیح می‌شوند.

۴-تکرار: برگشتن به مرحله دو و تکرار پروسه تا زمان همگرایی.

برای محاسبه وزن هر نورون در شبکه‌های عصبی از فرمول زیر استفاده می‌شود:

$$net_A = \sum_{i=0}^{n} w_{iA} x_i$$ **رابطه (۱)**

w_{iA} : وزن سیگنال ورودی نورون A و خروجی نورون i

x_i : وزن نورون i

n : تعداد نورون‌های لایه قبل

برای محاسبه میزان خروجی هر نورون نیز از فرمول زیر استفاده می‌شود:

$$output_A = \frac{1}{1 + e^{-x_A}}$$ **رابطه (۲)**

تا آخرین نورون برای تمام نورون‌ها این محاسبات انجام می‌شود. میزان خروجی بدست آمده برای نورون آخر را با مقدار خروجی واقعی مقایسه کرده و میزان دقت و خطا طبق فرمول زیر محاسبه می‌شود:

$$SSE = \sum_{records} \sum_{output\ nods} (Real\ Valu - Estimated\ Value)^2$$ **رابطه (۳)**

$Real\ Value$: مقدار خروجی واقعی

$Estimated\ Value$: مقدار خروجی محاسبه شده

حال وزن سیگنال‌های ورودی خروجی طبق فرمول‌های زیر از آخرین لایه تا اولین لایه تصحیح می‌شود:

$$W_{ij.\,new} = W_{ij.\,current} + \Delta W_{ij}$$ **رابطه (۴)**

$$\Delta W_{ij} = \eta \delta_j * output_i$$ **رابطه (۵)**

$$\delta_j = output_j(1 - output_j)(actual_j - output_j)$$ **رابطه (۶)**

η : مقداری ثابت که به صورت پیش فرض مشخص می‌شود.

$actual_j$: مقدار خروجی واقعی عصب j

برای لایه‌های میانی از فرمول زیر برای محاسبه δ_j استفاده می‌شود:

$$\delta_j = output_j(1 - output_j) \sum_{downstream} W_{jk.\,new}\ \delta_k$$ **رابطه (۷)**

۱/۴/۲. درخت‌های تصمیم

درخت تصمیم در حقیقت یک فلوچارت با ساختار درختی است. در این درخت برگ‌ها و گره‌هایی وجود دارند که این گره‌ها یا یک گره تصمیم هستند و یا یک گره برگ هستند. در هر گره تصمیم مقادیر یک متغیر بررسی می‌گردد و توسط شاخه‌ها انشعاب پیدا می‌کنند و در گره برگ نیز مقدار غالب متغیر رده رکوردهای موجود در آن انشعاب نوشته می‌شود.

برای تولید درخت تصمیم در حقیقت باید از شاخص‌هایی برای پیدا کردن خلوص انشعاب دهی استفاده نمود و متغیری را که خلوص بیشتری در انشعابات ایجاد می‌کند را در درخت استفاده کرد. در اینجا منظور از خلوص این است که پس از انشعاب در هر شاخه فقط رکوردهاییی که دارای یکی از رده‌های موجود در مجموعه داده هستند، حضور داشته باشند.

یکی از شاخص‌های انتخاب نقطه‌ی انشعاب ضریب جینی می‌باشد که به صورت زیر محاسبه می‌گردد:

$$I(s) = \frac{|s_1|}{|s|} I(s_1) + \frac{|s_2|}{|s|} I(s_2)$$ **رابطه(۸)**

$$I(s_1) = 1 - \sum p_i^2$$ **رابطه (۹)**

$|s|$: تعداد کل حالات موجود

$|s_1|$: تعداد کل حالات s_1

p_i : فراوانی نسبی برابر تعداد حالات هر کلاس تقسیم بر کل

انشعابات در این درخت یا دودویی هستند و یا چندگانه هستند[۱۴]. که درخت ایجاد شده در متلب برای این پژوهش یک درخت دودویی است.

۱/۵. مروری بر مطالعات انجام شده

در مطالعات انجام شده توسط اقای علی رفیعی و همکارانش با عنوان به کارگیری داده کاوی برای تشخیص تقلب در تراکنش کارت‌های اعتباری روش‌های مختلف داده کاوی مورد بررسی قرار گرفته است. در این مقاله الگوریتم‌های داده کاوی نظیر الگوریتم فازی، رگرسیون، درخت تصمیم، جنگل تصادفی، خوشه بندی، شبکه عصبی، برپایه قوانین، سیستم ایمنی مصنوعی، شبکه بیزین، روش‌های آماری بررسی شده است که برآیند مطالعه نشان می‌دهد استفاده از رهیافت‌های ترکیبی نظیر دو روش دسته بندی موجب بهبود نتایج از نظر معیارهایی نظیر دقت کشف، سرعت کشف ،هزینه کشف خواهد شد. به کارگیری مزیت‌های الگوریتم‌ها در کنار یکدیگر می‌تواند علت اصلی این بهبودها باشد[۵].

در مقاله انجام شده توسط خانوم زهره زرین قلمی تقلب در کارت‌های اعتباری از جمله تقلب ورشکستگی، کارت گمشده یا دزدی، تقلب درخواست، کنترل حساب، جعل کارت، تبانی فروشنده و تقلب‌های اینترنتی تشریح و سپس انواع روش‌های داده کاوی برای شناسایی تقلب در کارت‌های اعتباری مانند شبکه‌های عصبی، برپایه قوانین، درخت تصمیم، شبکه‌های بیزین، خوشه‌ای، آماری، الگوریتم‌های ژنتیک، منطق فازی رگرسیون، جنگل تصادفی بررسی شده است. به این نتیجه رسیده است که هیچ کدام از روش‌ها به تنهایی قادر به شناسایی تمامی انواع تقلب‌ها نیستند و هرکدام برای شناسایی نوع خاصی از تقلب مناسب هستند، به همین دلیل در بیشتر مراجع از چندین روش و نیز ترکیب آن‌ها استفاده شده و تنها در شمار بسیار کمی ازآن‌ها از یک روش به تنهایی استفاده شده است[۲].

در پژوهشی بهنام حیدری شبیه سازی در زمینه کشف تقلب کارت‌های اعتباری در سیستم‌های بانکداری الکترونیک انجام داده است. این پیاده سازی با استفاده از زبان برنامه نویسی C# و نرم افزارهای داده‌کاوی Rapidminer و Weka انجام شده است که دریافته استفاده از قوانین وابستگی در الگوریتم‌های دسته بندی مثل KNN می توان بهبود قابل توجهی در دقت دسته بندی ایجاد نماید. همچنین ترکیب تکنیک‌های خوشه بندی K-means با تعداد Kهای بهینه و قوانین انجمنی نقش بسیار مهمی در کشف تقلب داشته است. لذا دقت الگوریتم پیشنهادی به منظور کشف تقلب در داده‌های بانکداری الکترونیک، نسبت

به روش معمول بسیار مناسب بوده است. بنابراین در این مقاله یه روش جدید دسته بندی تلفیقی با خوشه بندی و قوانین انجمنی مبتنی بر قوانین وابستگی به منظور افزایش دقت کشف تقلب در سیستم‌های بانکداری الکترونیک ارائه شده است[۶].
هوانگ با استفاده از انتخاب ویژگی در مدل نظارت شده، مدل‌های آماری خطی و غیر خطی و مدل‌های یادگیری ماشین از قبیل شبکه عصبی، رگرسیون لاجستیک، درخت تقویت شده و جنگل تصادفی که بر مبنای داده‌های دارایی نیویورک و و داده‌های تراکنش کارت اعتباری تحقیق شده است به این نتیجه رسید مدل جنگل تصادفی بهترین عملکرد را دارد و می‌تواند نیمی از تلاش‌های مربوط به تقلب را تنها در ۳٪ داده‌های برتر به عنوان مشکوک با امتیاز الگوریتم تقلب تشخیص دهد. با این حال، نتایج خارج از زمان(out-of-time) جنگل تصادفی به دلیل مجموعه داده‌های کوچک بسیار نزدیک به شبکه عصبی است. از مزایای این تحقیق این است که برای مدل‌های خطی و غیر خطی استفاده می‌شوند و چون نظارت شده هستند، بسیار دقیق و قابل اعتمادند. نیاز به داده‌های بیشتر در فیلدهای بیشتر و همچنین مشکل داده‌های نا متوازن از معایب کار است[۱۰].
در پژوهش آقای دومان و همکارش تکنیک‌های پیشرفته داده‌کاوی به همراه الگوریتم شبکه عصبی برای پوشش بالای تقلب‌های انجام شده، معرفی شده است. به منظور مقایسه میزان قابلیت اطمینان و ودرصد صحت تشخیص تقلب الگوریتم‌های پیشنهادی، دو تکنیک شبکه عصبی و درخت تصمیم انتخاب شده‌اند و از آنجایی که هر دو روش انتخابی از نوع روش‌های یادگیری با ناظر می‌باشند، سعی شده تا این مقایسه از طریق تغییر در درصد یادگیری شان و سپس محاسبه درصد صحت تشخیص بر روی داده‌های مصنوعی، انجام شود. در ادامه به منظور دست یافتن به نتایج بهتر و قابل اعتمادتر، دو روش را با یکدیگر ادغام نموده و دوباره در میزان یادگیری‌های مختلف مقایسه صورت می‌گیرد. نتایج به خوبی نشان می‌دهد که ترکیب دو مدل در صورتیکه خروجی شبکه عصبی به عنوان منبع برای درخت تصمیم در نظر گرفته شود، در تشخیص تقلب کارت اعتباری کارآمد تر عمل می‌نماید[۸].

۲. روش پیشنهادی

برای کشف تقلبات صورت گرفته در تراکنش‌های کارت‌های اعتباری در این پژوهش به پیاده سازی الگوریتم‌های شبکه عصبی و درخت تصمیم در نرم افزار متلب پرداخته می‌شود و این دو الگوریتم را از نظر دقت پیش بینی مقایسه کرده و الگوریتم با خطای کمتر را معرفی می‌کنیم.
جهت اماده سازی داده برای استفاده از الگوریتم‌های یادگیری ماشین(درخت تصمیم و شبکه عصبی) ستون‌هایی از مجموعه داده که مقادیر آن‌ها عددی نیست را با استفاده از دستور grp2idx به مقادیر عددی تبدیل می‌کنیم چون الگوریتم یادگیری با مقادیر عددی کار می‌کند. سپس با استفاده از دستور normalize مقادیر ویژیگی‌ها را نرمالیزه می‌کنیم تا مقادیر آن‌ها در یک بازه استاندارد قرار بگیرند.
پس از شناخت و درک داده‌ها و مشخص نمودن فیلد هدف به پیاده سازی دو الگوریتم شبکه‌های عصبی و درخت تصمیم بر روی دو مجموعه داده معرفی شده می‌پردازیم. بدین صورت که در نرم افزار ذکر شده ابتدا مجموعه داده‌های موردنظر را بارگذاری کرده و برای شبکه عصبی در هر دو مجموعه داده ویژگی(Features) ازهمه ستون‌ها بجز ستون آخر استخراج می‌شود و بعنوان ورودی به الگوریتم داده می‌شود که دارای مجموعه‌ای از ویژگی‌های مشترک هستند و آخرین ستون به عنوان تارگت و هدف در نظر گرفته می‌شود که متقلب بودن یا نبودن را نشان می‌دهد. پس از این مرحله با استفاده از CVPartition ، مجموعه داده را به دو بخش مجموعه داده‌های آموزشی[۵] و مجموعه تست[۶] تقسیم کرده که با استفاده از مجموعه داده آموزشی، الگوریتم آموزش می‌بیند و به یادگیری می‌پردازد و مدل نهایی ایجاد می‌شود و سپس با استفاده از مجموعه تست، مدل ایجاد شده مورد ارزیابی و اعتبارسنجی قرار می‌گیرد. یک شبکه عصبی با سه لایه پنهان که هر یک شامل ۱۰ نورون

[5] Training Set
[6] Test Set

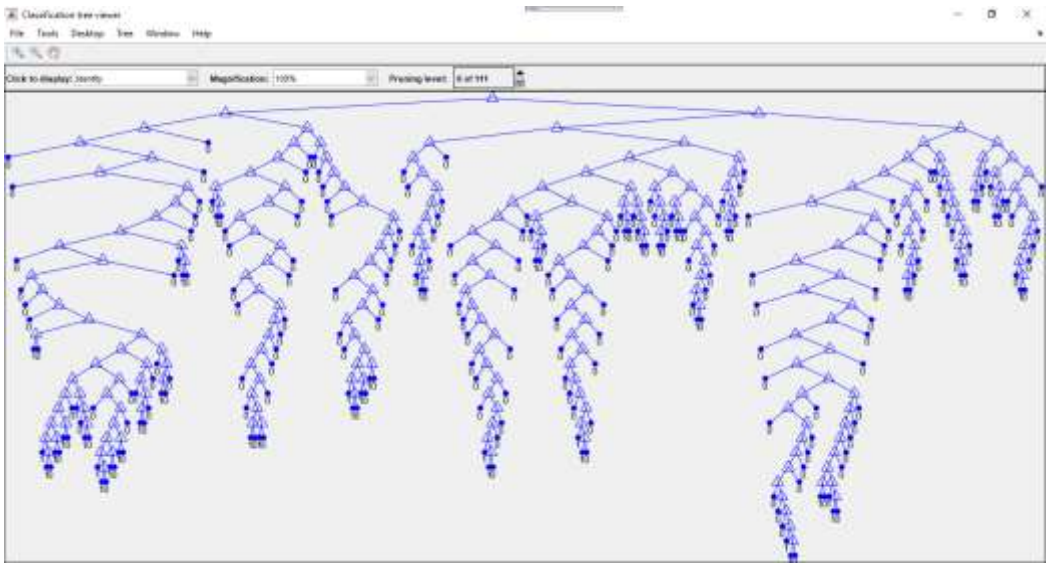
است، با استفاده از patternnet تعریف می‌شود و الگوریتم شبکه عصبی روی مجموعه داده‌ها اعمال می‌شود. قابل ذکر است که این تعداد لایه و نورون به این دلیل انتخاب شده که الگوریتم دقت خوب و بالایی داشته باشد. اگر تعداد نورون و لایه کم باشد، نتیجه با دقت بالایی حاصل نمی‌شود و از طرفی اگر تعداد نورون و لایه‌ها زیاد باشد، حجم محاسبات و زمان ران بالا می‌رود.
در شکل ۱ شبکه عصبی ساخته شده با ساختار ذکر شده نشان داده شده است.

شکل۱ شبکه عصبی ساخته شده توسط الگوریتم

برای درخت تصمیم نیز مانند شبکه عصبی در هر دو مجموعه داده ستون هدف یا تارگت مشخص می‌شود و داده‌ها با استفاده از CVPartition مجموعه داده را به دو بخش داده‌های آموزشی و داده‌های آزمایش تقسیم بندی می‌کنیم. قبل از ایجاد درخت تصمیم‌گیری این نکته را باید در نظر داشته باشیم که ما ابتدا باید نمونه‌های مثبت و منفی را برای مدیریت عدم تعادل کلاس در مجموعه داده را شناسایی کنیم، نمونه‌های مثبت برچسب "۱" و نمونه‌های منفی برچسب "۰" دارند. سپس از کلاس اقلیت(نمونه‌های مثبت) نمونه برداری بیش از حد انجام دادیم(بیش نمونه برداری یا OverSample کردیم) تا تعداد نمونه‌های مثبت و منفی را متعادل کنیم و برای این کار از نمونه برداری با جایگزینی استفاده می‌کنیم. این کار برای جلوگیری از سوگیری درخت تصمیم به سمت کلاس اکثریت(نمونه‌های منفی) انجام می‌شود و اطمینان حاصل بشود که مدل به نفع کلاس اکثریت نیست در غیر اینصورت می‌تواند منجر به عملکرد ضعیف در پیش بینی کلاس اقلیت شود. برای شبکه عصبی نمونه برداری بیش از حد یا oversample را اجرا نکردیم چون شبکه عصبی به دلیل پیچیدگی و توانایی یادگیری الگوهای پیچیده تر از داده‌ها و مکانیسم‌های متفاوتی مانند backpropagation و معماری‌های پیچیده تر، ذاتا می‌تواند عدم تعادل کلاس را بهتر و موثرتر مدیریت کند. پارامترهای استفاده شده در ساخت درخت تصمیم دودویی در متلب به شرح زیر می باشد:

- MaxNumSplts: حداکثر تعداد انشعابات در درخت می‌باشد که در اینجا ۵۰۰۰ در نظر گرفته شده است تا درخت تصمیم بتواند عمیقا رشد کند و الگوهای پیچیده تری را در داده‌ها ثبت کند. این زمانی مفید است که داده‌ها پیچیده هستند و برای دسته بندی دقیق نیاز به مدلی دقیق دارند.
- MinLeafSize: حداقل اندازه هر برگ می‌باشد که یک در نظر گرفته شده است که درخت می‌تواند به اندازه لازم برگ‌های کوچک ایجاد کند و در صورت نیاز الگوهای بسیار خاصی را ثبت کند. این مورد می واند به بهبود دقت کمک کند، اما خطر بیش از حد پردازش را نیز افزایش می‌دهد.
- SplitCriterion: معیار تقسیم بندی می‌باشد که gdi در نظر گرفته شده است. (GDI) شاخص تنوع جینی معیاری از ناخالصی است که برای تقسیم گره‌ها استفاده می‌شود. هدف آن ایجاد گره‌هایی با کلاس‌های خالص یا همگن است.
- PredictorSelection: روش انتخاب پیش بینی کننده می‌باشد که curvature در نظر گرفته شده است. این گزینه از انحنا برای انتخاب بهترین پیش بینی کننده‌ها برای تقسیم استفاده می‌کند و به انتخاب ویژگی‌هایی که رابطه غیر خطی با متغیر هدف دارند کمک می‌کند.
- Surrogate: استفاده از متغیرهای جایگزین می‌باشد که این گزینه را در حالت on قرار دادیم. گزینه on یا روشن برای جانشین‌ها به درخت اجازه می‌دهد از تقسیم‌های جایگزین استفاده کند که می‌تواند با ارائه معیارهای تقسیم جایگزین، داده‌های از دست رفته را مدیریت کند.

در شکل ۲ بعنوان نمونه درخت تصمیم ساخته شده توسط الگوریتم برای مجموعه داده Creditcard با درصد داده اموزشی ۸۰ درصد و داده آزمایشی ۲۰ درصد نشان داده شده است.



شکل ۲ درخت تصمیم ساخته شده توسط الگوریتم برای دیتاست Creditcard با train=80%

در این پژوهش در دو مرحله الگوریتم‌های مورد نظر مورد بررسی قرار گرفتند، به طوریکه در مرحله اول ۸۰ درصد از داده‌ها به عنوان داده‌های آموزشی و ۲۰ درصد دیگر به عنوان مجموعه تست تعیین شد و مدل نهایی ایجاد و مورد بررسی قرار خواهد گرفت. در مرحله بعد ۷۰ درصد از داده‌ها به عنوان داده‌های آموزشی و ۳۰ درصد دیگر به عنوان مجموعه تست تعیین شد و مدل مورد بررسی قرار می‌گیرد.

۳. ارزیابی نتایج

برای پیاده سازی الگوریتم‌های درخت تصمیم و شبکه عصبی بر روی دو مجموعه داده که در ادامه به معرفی آن‌ها می‌پردازیم، از نرم افزار قدرتمند متلب نسخه ۲۰۲۳ استفاده شده که اجرا شده بر روی سیستم با ویندوز 10 Pro و پردازنده Core i5 و رم 4GB می‌باشد.

همچنین برای اجرای این دو الگوریتم از جعبه ابزار(ToolBax) آماده متلب و توابع موجود در آن استفاده شده است. برخی از توابع استفاده شده شامل تابع readtable برای خواندن مجموعه داده، تابع cvpartition برای تقسیم بندی داده ها به دو بخش آموزش و تست، تابع patternnet برای ایجاد شبکه عصبی، تابع fitctree برای ایجاد درخت تصمیم، تابع datasample برای oversample یا نمونه برداری بیش از حد، تابع train برای آموزش الگوریتم و تابع confusion برای ایجاد ماتریس آشفتگی می‌باشد.

از دو مجموعه داده که از سایت کاگل تهیه شده در این پژوهش استفاده شده است. مجموعه داده اول به نام Creditcard که شامل معاملات انجام شده توسط کارت‌های اعتباری در سپتامبر ۲۰۱۳ می‌باشد که معاملات رخ داده طی دو روز می‌باشد و از بین ۲۸۴۸۰۷ معامله ۴۹۲ معامله متقلبانه بوده. این مجموعه شامل تنها متغیرهای عددی می‌باشد و با توجه به مسائل مربوط به محرمانگی اطلاعات ویژگی‌های اصلی و اطلاعات پس زمینه بیشتر ارائه نشده است. این مجموعه داده دارای ۲۹ ستون ویژگی و یک ستون هدف می باشد که متقلبانه بودن یا نبودن تراکنش را مشخص میکند. مجموعه داده دوم به نام Banksim می باشد که این مجموعه داده به صورت مصنوعی تولید شده که شامل پرداخت‌هایی از مشتریان مختلف است که در دوره‌های زمانی مختلف و با مبالغ مختلف انجام می‌شوند. مجموعه داده دارای ۹ ستون ویژگی و یک ستون هدف است.

۳/۱. معرفی معیارهای ارزیابی

برای اعتبارسنجی و مقایسه کارائی الگوریتم‌ها از معیارهای مختلفی استفاده می‌شود. از جمله این معیارها میتوان به نرخ فراخوانی[۷]، دقت[۸]، صحت[۹] و F-mesure اشاره نمود که پس از تشکیل ماتریس پراکندگی یا درهم ریختگی متقلب و غیرمتقلب که در جدول ۱ نشان داده شده است به راحتی قابل محاسبه هستند. پارامترهای عنوان شده در این ماتریس به شرح ذیل می‌باشد[۱۵].

پارامتر (TN)[۱۰]: بیانگر تعداد رکوردهایی که دسته واقعی آن‌ها متقلب بوده و الگوریتم دسته بندی نیز دسته آن‌ها را به درستی متقلب پیش بینی کرده.

پارامتر (FN)[۱۱]: بیانگر تعداد رکوردهایی که دسته واقعی آن‌ها غیرمتقلب بوده و الگوریتم دسته بندی دسته آن‌ها را به اشتباه متقلب پیش بینی کرده.

پارامتر (TP)[۱۲]: بیانگر تعداد رکوردهایی که دسته واقعی آن‌ها غیرمتقلب بوده و الگوریتم دسته بندی دسته آن‌ها را به درستی غیر متقلب پیش بینی کرده.

پارامتر (FP)[۱۳]: بیانگر تعداد رکوردهایی که دسته واقعی آن‌ها متقلب بوده و الگوریتم دسته بندی دسته آن‌ها را به اشتباه غیرمتقلب پیش بینی کرده.

[7] Recall
[8] Precision
[9] Accuracy
[10] True Negative
[11] False Negative
[12] True Positive
[13] False Positive

جدول ۱ - ماتریس رفتارهای متقلبانه و غیر متقلبانه

نوع رکورد	رکوردهای پیش بینی شده(Predicated Records)		
	نوع دسته	غیر متقلب	متقلب
رکوردهای واقعی (Actual Records)	غیرمتقلب	TP	FN
	متقلب	FP	TN

مهم ترین معیار برای تعیین کارایی یک الگوریتم دسته بندی، معیار صحت می‌باشد، این معیار دقت کل یک دسته بند را محاسبه می‌کند. این معیار نشان دهنده این موضوع است که چند درصد از کل مجموعه داده‌ها به درستی دسته بندی شده است رابطه زیر نحوه محاسبه معیار را نشان می‌دهد[۱۵].

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$ **رابطه (۱۰)**

دو مقدار TN و TP مهم‌ترین مقادیری هستند که باید بیشینه شوند تا کارایی دسته بندی به حداکثر برسد. معیارهای دقت درصدی را نشان می‌دهند که از میان تمامی دسته‌ها که توسط دسته بند به آن دسته نسبت داده شده‌اند، درست دسته بندی شده‌اند. نحوه محاسبه این معیار در رابطه زیر نشان داده شده است.

$$Precision = \frac{TP}{TP + FP}$$ **رابطه(۱۱)**

معیار فراخوانی برای یک دسته، که از میان تمامی دسته‌های متقلب متعلق به آن دسته، به درستی دسته بندی شده است. نحوه محاسبه این معیار به صورت زیر می‌باشد.

$$Recall = \frac{TP}{TP + FN}$$ **رابطه (۱۲)**

معیار F-measure از ترکیب معیارهای Precision و Recall بدست می‌آید و در مواردی استفاده می‌شود که نتوان اهمیت ویژه‌ای را برای هر یک از دو معیار Precision و Recall نسبت به یکدیگر قائل شد. رابطه زیر نحوه محاسبه این معیار را نشان می‌دهد[۷].

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall}$$ **رابطه (۱۳)**

برای ارزیابی روش‌های دسته بندی از معیارهای بالا استفاده خواهد شد.

۳/۲. ارزیابی

هدف از دسته بندی، ایجاد مدلی بر مبنای داده‌های آموزشی عنوان شده که بتواند صفت کلاس را برای داده‌های مجموعه تست پیشگویی کند. همان طور که گفته شد مجموعه تست برای ارزیابی و اعتبارسنجی مدل ساخته شده مورد استفاده قرار می‌گیرد به همین دلیل مقادیر گزارش شده برای معیارهای ارزیابی در جداول و نمودارها، بدست آمده از نتایج مجموعه تست مدل ساخته شده می‌باشد. برای مقایسه کارایی الگوریتم‌ها این نتایج در جداول و روی نمودارها قرار داده شده است و در ادامه به تشریح آن‌ها پرداخته می‌شود.

در جداول ۲، ۳، ۴ و ۵ نتایج بدست آمده از مرحله اول و دوم پیاده سازی قرار داده شده است. که در مرحله اول ۸۰ درصد از داده‌ها به عنوان مجموعه آموزشی و ۲۰ درصد دیگر به عنوان مجموعه تست در نظر گرفته شده بود و در مرحله دوم ۷۰ درصد از داده‌ها به عنوان مجموعه آموزشی و ۳۰ درصد دیگر به عنوان مجموعه تست در نظر گرفته شد. در سطر دوم و سوم جداول به ترتیب، مقادیر معیارهای ارزیابی برای الگوریتم شبکه عصبی و الگوریتم درخت تصمیم قرار داده شده است. در ستون دوم جدول مقادیر معیار Precision که بیان کننده دقت الگوریتم در پیش بینی می‌باشد برای هر دو الگوریتم قرار گرفته است؛ در ستون سوم مقادیر معیار Recall که بیان کننده درصد فراخوانی هر الگوریتم می‌باشد قرار گرفته است؛ در ستون سوم مقادیر معیار Accuracy که بیان کننده درصد صحت پیش بینی انجام شده توسط الگوریتم‌ها می‌باشد قرار گرفته است و در ستون چهارم نیز مقادیر معیار F-measure برای هر دو الگوریتم قرار گرفته است.

جدول۲ نتایج بدست آمده برای دیتا ست Creditcard با train=80% و Test=20%

معیار(درصد) / الگوریتم	Precision	Recall	Accuracy	F-measure
Neural Net	۹۹.۹۸۵۹	۹۹.۹۰۸۶	۹۹.۸۹۴۷	۹۹.۹۴۷۲
Decision Tree	۹۹.۹۶۴۷	۱۰۰	۹۹.۹۸۲۴	۹۹.۹۸۲۴

جدول Error! No text of specified style in document. **نتایج بدست آمده برای دیتا ست Creditcard با train=70% و Test=30%**

معیار(درصد) / الگوریتم	Precision	Recall	Accuracy	F-measure
Neural Net	۹۹.۹۶۷۲	۹۹.۹۶۰۱	۹۹.۹۲۷۴	۹۹.۹۶۳۷
Decision Tree	۹۹.۹۵۳۲	۱۰۰	۹۹.۹۷۶۶	۹۹.۹۷۶۶

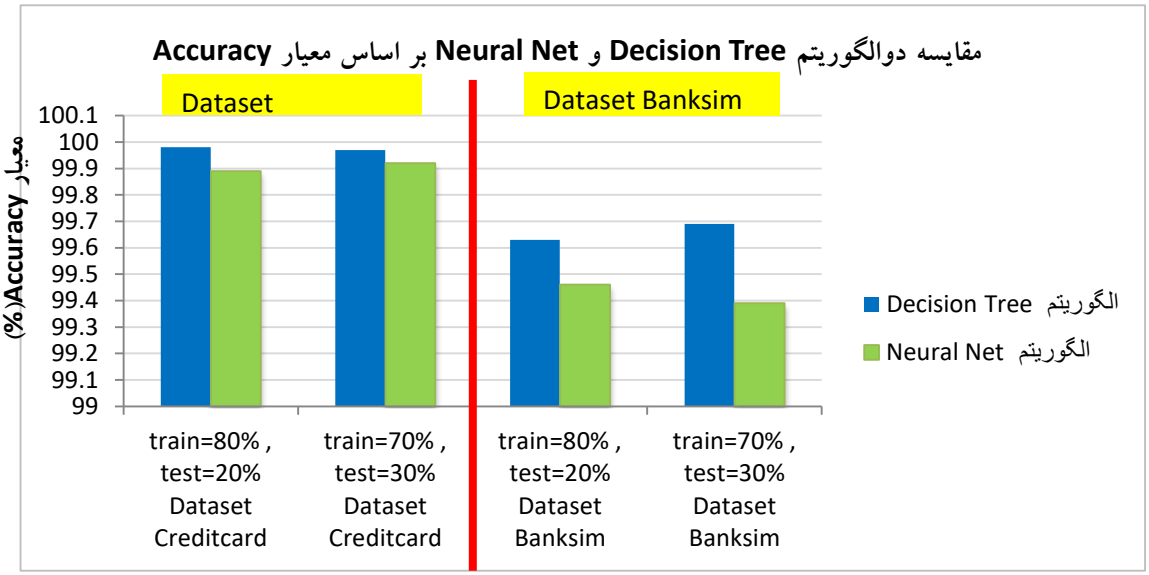
جدول Error! No text of specified style in document. **نتایج بدست آمده برای دیتا ست Banksim با train=80% و Test=20%**

معیار(درصد) / الگوریتم	Precision	Recall	Accuracy	F-measure
Neural Net	۹۹.۹۲۰۹	۹۹.۵۳۶۳	۹۹.۴۶۱۹	۹۹.۷۲۸۲
Decision Tree	۹۹.۲۷۸۷	۱۰۰	۹۹.۶۳۴۵	۹۹.۶۳۸۱

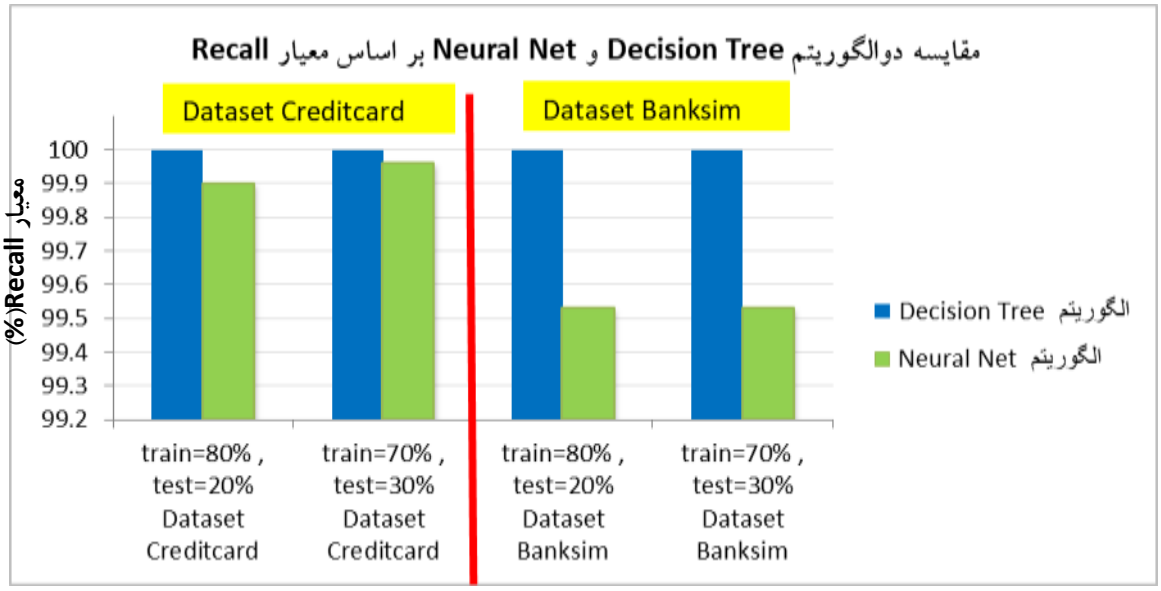
جدول Error! No text of specified style in document. **نتایج بدست آمده برای دیتا ست Banksim با train=70% و Test=30%**

معیار(درصد) / الگوریتم	Precision	Recall	Accuracy	F-measure
Neural Net	۹۹.۸۶۲۷	۹۹.۵۳۰۷	۹۹.۳۹۹۱	۹۹.۶۹۶۴
Decision Tree	۹۹.۴۰۲۹	۱۰۰	۹۹.۶۹۷۱	۹۹.۷۰۰۶

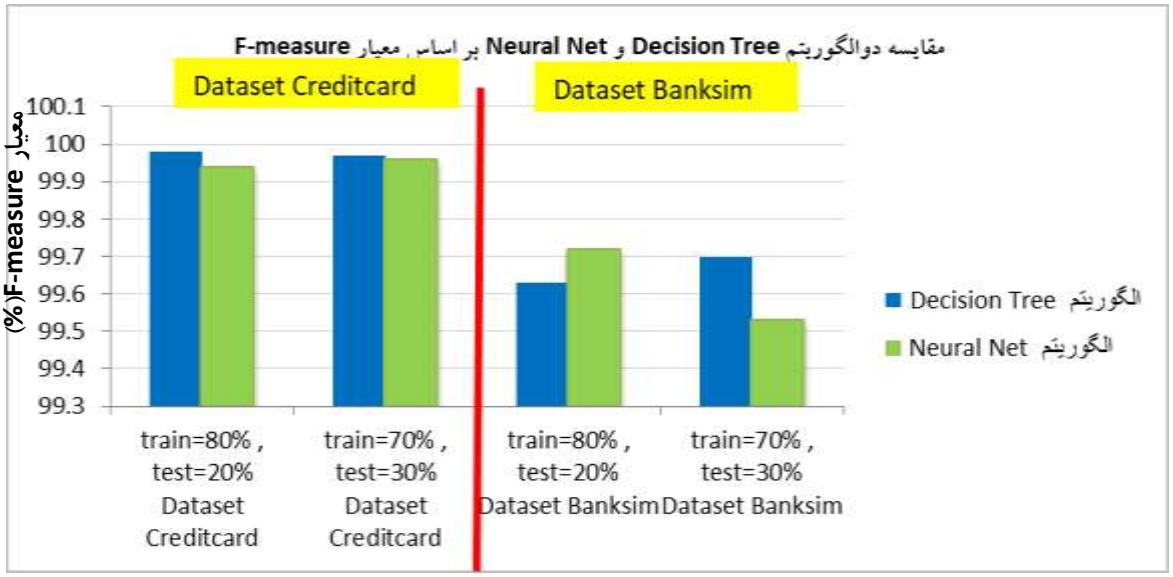
همانطور که مشاهده می‌کنید طبق نتایج به دست آمده در جداول ۲، ۳، ۴ و ۵ مقادیر Precision برای الگوریتم شبکه عصبی کمی بیشتراز درخت تصمیم بوده است ولی در هر دو الگوریتم این مقادیر درصد مناسبی برای دقت الگوریتم‌ها می‌باشد ولی برای مقایسه نمی‌توان از آن استفاده کرد. مقدار F-measure نیز در هر مرحله مقدار قابل توجهی داشته و در حالت کلی درخت تصمیم در هر مرحله درصد بیشتری اخذ کرده است.معیار Recall در هر مرحله برای الگوریتم درخت تصمیم مقدار برابر ۱۰۰ درصد را اخذ کرده و نسبت به الگوریتم شبکه عصبی درصد بالاتری را به خود اختصاص داده است.و اما معیار Accuracy همانطور که اشاره شد مهم‌ترین معیار برای الگوریتم‌های دسته بندی می‌باشد که این معیاردر همه مراحل، مقادیر الگوریتم درخت تصمیم نسبت به الگوریتم شبکه عصبی درصد بیشتری را اخذ نموده و عملکرد بهتری داشته است.مقایسه عملکرد دو الگوریتم Neural Net و Decision Tree بر اساس سه معیار Accuracy، Recall و F-measure روی نمودار در شکل های ۳، ۴ و ۵ نشان داده شده است.



شکل Error! No text of specified style in document. **مقایسه دوالگوریتم Decision Tree و Neural Net بر اساس معیار Accuracy**



شکل Error! No text of specified style in document. **مقایسه دوالگوریتم Decision Tree و Neural Net بر اساس معیار Recall**



شکل Error! No text of specified style in document. **مقایسه دوالگوریتم Decision Tree و Neural Net بر اساس معیار F-measure**

به طور کلی همانطور که در نمودارها نیز مشخص است، در هر مرحله از پیاده سازی با توجه به معیارهای Accuracy، Recall و F-measure، الگوریتم درخت تصمیم عملکرد بهتری نسبت به الگوریتم شبکه عصبی داشته است.

۴. بحث و نتیجه گیری

برای تحلیل تراکنش‌های انجام شده در کارت‌های اعتباری توسط مشتری که از عامل‌های مهم و تاثیرگذار در کشف تقلب است، از الگوریتم‌های دسته بندی درخت تصمیم و شبکه عصبی که به بیان دانش و استخراج در ارتباط با مجموعه‌ای از اطلاعات می‌پردازد، استفاده شده است. طبق گزارشات صورت گرفته و نتایجی که از دو الگوریتم انتخابی طی مراحل پیاده‌سازی بدست آمد؛ الگوریتم درخت تصمیم در هر مرحله عملکرد بهتری نسبت به الگوریتم شبکه‌های عصبی داشت اگر چه طبق معیارهای ارزیابی بدست آمده هر دو الگوریتم درصد قابل توجهی از معیارها را جهت پیش‌بینی و کشف تقلبات به خود اختصاص دادند. در آخر برای جمع‌بندی و بهبود دستاورد پژوهشی مقایسه‌ای با دیگر پژوهش‌های ارائه شده در این زمینه انجام دادیم که طبق اکثر پژوهش‌های انجام شده، پژوهشگران به این نتیجه مهم دست یافتند که هیچ کدام از روش‌ها به تنهایی قادر به شناسایی تمامی انواع تقلب‌ها نیستند و هر کدام برای شناسایی نوع خاصی از تقلب مناسب هستند بنابراین تکنیک‌های داده‌کاوی به تنهایی کارایی بالایی ندارند و بهتر است از رویکرد ترکیبی آن‌ها استفاده شود به همین دلیل در بیشتر مراجع از چندین روش و نیز ترکیب آن‌ها استفاده شده است و تنها در شمار بسیار کمی از آن‌ها از یک روش به تنهایی استفاده شده است.

۵. منابع

۱-سیادت، سیدحسین؛ رباب اسمعیلی، هما؛ عزیزی، طوفان؛ "بررسی و مقایسه روش های داده کاوی جهت کشف تقلب در تراکنش های کارت های بانکی"، دومین کنفرانس بین المللی مدیریت در قرن ۲۱، ۱۳۹۴.

۲-زرین قلمی، زهره؛ "به کارگیری داده کاوی برای تشخیص تقلب در تراکنش های کارت های اعتباری"، دومین همایش ملی فناوری های نوین در مهندسی برق و کامپیوتر، دانشگاه آزاد اسلامی فسا، ۱۳۹۳.

۳-فصیخ فر، زهره؛ رخصتی، حمیدرضا؛ وفاخواه، محبوبه؛ "بررسی کاربرد الگوریتم های هوشمند داده کاوی، در شناسایی و جلوگیری از انواع تقلبات بانکی"، سومین کنفرانس بین المللی مهندسی کامپوتر و سیستم های خبره، ۱۳۹۵.

۴-صنیعی آباده محمد، محمودی سینا، طاهرپور محدثه، "داده کاوی کاربردی"، تهران ، نیاز دانش، ۱۳۹۱.

۵-رفیعی کشتلی، علی؛ رفیعی کشتلی، میثم؛ دکتر شمسی، محبوبه "به کارگیری داده کاوی برای تشخیص تقلب در تراکنش کارت های اعتباری"؛ اواین همایش ملی؛ ۱۳۹۲.

۶-حیدری، بهنام؛ " ارائه یک روش بهینه جهت بهبود کشف تقلب های کارت اعتباری در سیستم بانکداری الکترونیک با استفاده از ترکیب رویکردهای داده کاوی"؛ چهارمین کنفرانس بین المللی مهندسی برق و کامپیوتر؛ ۱۳۹۵.

۷-نجف زاده اصل، احسان و کریم پور، جابر و مقدم دیزج هریک، مجید؛ "ارائه‌ی یک روش جدید کشف نفوذ به منظور تشخیص رفتارهای غیرنرمال در شبکه های کامپیوتری با استفاده از تکنیک‌های داده کاوی"، چهاردهمین کنفرانس مهندسی برق ایران، کرمانشاه، ۱۳۹۰.

8-Zakaryazad, Ashkan, & Ekrem Duman, "A profit-driven Artificial Neural Network(ANN) with applications to fraud detection and direct marketing" Neurocomputing 175, 2016.

9-Zhang, Zhaohui, et al. "A model based on convolutional neural network for online transaction fraud detection", Security and Communication Networks, 2018.

10-Huang, "Credit Card Transaction Fraud Using Machine Learning Algorithms, International Conference on Education science and Economic Development, 2020.

11-https://en.wikipedia.org/wiki/cretid_card_fraud.

12-El Madhoun, Nour, and Emmanuel Bertin. "Magic always comes with a price: Utility versus security for bank cards." 2017 1st Cyber Security in Networking Conference (CSNet). IEEE,2017.

13-Miner, G., Nisbet, R., IV, J.E.; "Handbook of Statistical Analysis and Data Mining Applications, 1 edition ed", *Academic press, Amsterdam, Boston*, 2009.

14-Witten, I. H., & Frank, E.; "Data Mining: Practical machine learning tools and techiques", Morgan Kaufmann, 2005.

15-Raval, U. Marathe, N. Padiya, P.; "Feature Selection Based Hybrid Anomaly Intrusion Detection System Using K Means and RBF Kernel Function", scientific committee of International Conference on Advance

Computing Technologies and Application (ICACTA-2015), Published by Elsevier B.V. Procedia Coputer Science 45, PP: 428-435, 2015.

Fraud detection in credit cards using neural network and decision tree

Reza Roshani
1-Lamei Gorgani Institue of Higher Education, Gorgan, Iran
2-Department of Mechanical Engineering, National University of Skills (NUS), Tehran, Iran

Masoumeh Kiaee
1-Lamei Gorgani Institue of Higher Education, Gorgan, Iran

Abstract:

Using credit cards to prevent accidents caused by carrying cash has become popular. But this issue not only has not caused the retreat of profiteers, but has opened new ways with less risk for them. From the analysis of bank statistics and other credit card custodians, it appears that fraud in credit cards is increasingly increasing, and due to the high volume of transactions, the need for a suitable tool to detect data fraud is felt. Data mining is a new science that has been proposed in this field, so that by using its techniques and algorithms, it is possible to process this massive amount of data and search and identify suspicious cases. In this research, in order to provide a method in the field of fraud detection in credit card transactions, two widely used data mining techniques, namely, decision tree and neural network, are investigated on two data sets, and these algorithms are analyzed in two stages using MATLAB data mining tool. They will be implemented on selected data sets and their results will be analyzed and compared. The implementation results show that the decision tree algorithm has a higher percentage of the criteria than the neural network.

Keywords: Credit card, fraud detection, data mining, decision tree, neural network.