همایش ملی الکترونیکی دستاوردهای نوین در علوم مهندسی و پایه
National e-Conference on Advances in Basic Sciences and Engineering
. . . . WWW.AEBSCONF.IR . . . .

کاربرد داده‌کاوی جهت کشف رفتارهای مشکوک در تراکنش های‌کارت‌های بانکی

[۱] لیلا جدیدی، [۲] پروین بهاران

۱- عضو هیأت علمی دانشگاه آزاد اهواز مؤسسه فرهنگی آموزشی سماء خوزستان، اهواز، ایران
۲- دانشجوی کارشناسی نرم افزار مؤسسه غیرانتفاعی جهاد دانشگاهی خوزستان، اهواز، ایران
Leila jadidi@yahoo.com

چکیده

تکنیک‌های داده‌کاوی به صورت گسترده ای در دنیا در حوزه کشف تقلب مورد استفاده قرار گرفته شده‌است، دلیل این موضوع انطباق زیاد این فرآیند با نیازمندی‌های مساله کشف تقلب می‌باشد. در این مقاله سعی بر پیاده‌سازی قدم به قدم فرآیند داده‌کاوی بر داده های مربوط به تراکنش‌های کارت بانکی فردی خاص، جهت کشف رفتار مشکوک در این تراکنش‌ها شده‌است. در مرحله مدل‌سازی این فرآیند از تکنیک تحلیل خوشه جهت آموزش مدل‌هایی با پارامترها و تنظیمات متنوع (برای مثال تعداد خوشه‌های مختلف، تعداد نرون‌های مختلف و...) بر داده های بدون برچسب، استفاده شده‌است. تکنیک خوشه بندی انتخاب گردیده در این مقاله نگاشت خود سازمانده یا همان شبکه عصبی کوهنن می‌باشد، بعد از خوشه‌بندی داده ها از مدل طبقه‌بندی، داده ها از مدل طبقه‌بندی، جهت ارزیابی مدل‌های غیر نظارتی و همچنین انتخاب بهینه‌ترین مجموعه مدل ترکیبی (مدل‌های نظارتی و غیرنظارتی) جهت کشف رفتار مشکوک تراکنش‌های آتی، استفاده شده‌است. تکنیک طبقه‌بندی انتخاب شده، مدل مجموعه قوانین با استفاده از الگوریتم C5 می‌باشد. در انتهای تحقیق نیز، قوائد مدل طبقه‌بندی بهینه، جهت استخراج دانش‌هایی مورد تفسیر قرار گرفته است.

واژه های کلیدی: داده‌کاوی، رفتار مشکوک، مدل‌های غیر نظارتی، تحلیل خوشه، شبکه عصبی کوهنن

مقدمه

افزایش تنوع کارت‌های اعتباری و بانکی ارائه شده از سوی بانک‌ها و موسسات مالی و اختصاص سهم بالایی از حجم مبادلات مالی افراد حقیقی و حقوقی در دنیا توسط این کارت‌ها، احتمال ایجاد تخلف و سوء استفاده از این کارت‌ها و ایجاد مبالغ بالای ضرر و زیان نیز برای مشتریان و موسسات مالی رو به افزایش می‌دهد، برای مثال لئونارد هزینه‌های مربوط به تقلب برای کارت‌های ویزا و مستر کارت در کانادا را به ترتیب برای سال‌های ۱۹۸۹ و ۱۹۹۰ و ۱۹۹۱ مبالغی معادل با ۱۹ و ۲۹ و ۴۶ میلیون دلار برآورد نموده است. همچنین منبع آماری اپکس روند رشد تقلب کارت‌های اعتباری در کشور انگلستان در سال‌های ۱۱۹۷ و ۲۰۰۰۰ را به ترتیب برابر با مبالغ ۱۲۲ و ۱۳۵ و ۱۸۸ و ۲۹۳ میلیون یورو برآورد نموده است.[۱]

بیشتر مسائل کشف تقلب در حوزه کارت‌های بانک، شامل حجم عظیمی از مجموعه داده می‌باشد، برای مثال شرکت کارت‌های اعتباری بارکلای کارد تقریبا به تنهایی دارای ۳۵۰ میلیون تراکنش در سال در کشور انگلستان می‌باشد. به دلیل حجم بالای داده در پایگاه اطلاعاتی مربوط به تراکنش‌های کارت‌های بانکی و همچنین وجود اطلاعاتی بصورت عددی و اسمی (سمبولیک) و یا ترکیبی از این دو، پرداز بر روی این مجموعه داده ها برای جستجو تراکنش های متقلبانه، نیاز به استفاده از الگوریتم‌های کارا و سریع دارد. با معرفی و رشد دانش داده‌کاوی در دنیا و همچنین برآورده نمودن نیازمندی‌های مساله کشف تقلب، یعنی امکان استفاده از این ابزار بر روی حجم بالای اطلاعات و کشف الگوهای پنهان در این داده ها بصورت خودکار و سریع، شاهد توسعه بکارگیری این رویکرد در حل مسائل کشف تقلب در حوزه کارت‌های بانکی در سال‌های اخیر در دنیا بوده‌ایم.

تقسیم‌بندی مدل‌های کشف تقلب در حوزه کارت‌های بانکی

استراتژی‌های کلان مسائب کشف تقلب در حوزه کارت‌های بانکی را نیز می‌توان، منطبق با استراتژی‌های فرآیند داده‌کاوی دانست.

دو استراتژی کلان برای فرآیند داده‌کاوی قابل تصور می‌باشد که عبارتند از:

یادگیری نظارتی (هدایت شده)

یادگیری غیرنظارتی (غیر هدایتی)

انتخاب یکی از این دو رویکرد و استفاده از مدل‌ها و روش‌های قابل استفاده مرتبط با آن‌ها نیز بستگی به ماهیت اطلاعات و داده‌های تاریخی موجود دارد. زمانی که داده های موجود بصورت برچسب خورده باشند می‌توان از رویکر نظارتی استفاده نمود و در غیر این‌صورت رویکرد غیر نظارتی (هدایت نشده) رویکرد مناسب تری جهت این موضوع می‌باشد. تکنیک‌ها و مدل‌های مورد استفاده در خصوص مساله کشف تقلب در سال ۲۰۰۲ توسط بلتن و هند بررسی مروری گردید، آن‌ها مدل‌های کشف تقلب در حوزه کارت‌های اعتباری را تحت دو رویکرد کلی نظارتی و غیر نظارتی طبقه‌بندی نمودند.[۱] همچنین فوا، لی، اسمیت و گایلر نیز در سال ۲۰۰۵ تکنیک های مرتبط با حوزه کارت‌های اعتباری را در چهار گروه ۱- گروه‌های نظارتی ۲ – غیر نظارتی ۳ – ترکیب تکنیک‌های نظارتی ۴ – ترکیب تکنیک‌های نظارتی تقسیم‌بندی نمودند.[۲]

بیان مساله و رویکرد اصلی تحقیق

امکان انجام پروژه های داده‌کاوی مستلزم استفاده از داده های مرتبط در خصوص مساله می باشد، به همین منظور داده های تراکنش‌های کارت‌های بانکی به عنوان اصلی‌ترین ورودی مدل‌های کشف تقلب، از یکی از بانک‌های موجود در کشور برای انجام این تحقیق جمع‌آوری شده است و از آنجا که امکان دسته‌بندی داده های مربوط به تراکنش‌های ...یگر بر چستب زنی داده ها) در بیشتر بانکها و موسسات مالی و از جمله بانک مورد نظر، ...ز متداول‌ترین متدولوژی های داده کاوی به ویژه استفاده از مدلهای غیر نظارتی، بر داده

آکادمی داده
dataacademy.ir

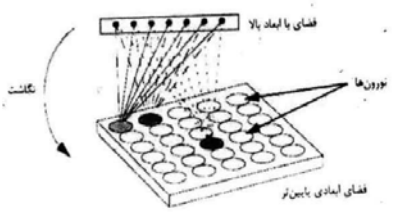
های تراکنش‌های کارت‌های بانکی مربوط به فردی خاص در این بانک می‌باشد، این امر به جهت طراحی سرویسی ویژه در بانک مورد نظر برای مشتریان بانک در خصوص امکان ایجاد کنترل و نظارت بر تراکنش‌های صورت گرفته از طریق کارت‌های متعلق به آن‌ها می‌باشد.

مرور ادبیات مرتبط با تحقیق

تکنیک مورد استفاده در این تحقیق نگاشت خود سازمانده (شبکه عصبی کوهنن) جهت تحلیل خوشه داده ها با رویکرد غیرنظارتی می‌باشد. در سال ۲۰۰۶ زاسلاویسکی و استریژاک اقدام به ارائه مدلی برگرفته از الگوریتم نگاشت خود سازمانده جهت تشخیص تراکنش‌های مشکوک کارت‌های اعتباری با رویکرد غیرنظارتی، برای سیستم پرداخت بانک‌ها نمودند.[۳] همچنین کواه و اسریگانش در سال ۲۰۰۸ با توسعه رویکرد قبلی و با تاکید بر حل مساله کشف تقلب در زمان حقیقی با حداقل هزینه و زمان، اقدام به طراحی مدلی ب سه لایه نمودند. لایه اصلی و هسته‌ای ابتدا با استفاده از نگاشت خود سازمانده اقدام به غربالگری و دسته بندی اولیه داده ها به دو گروه تراکنش‌های مشکوک قانونی نموده و سپس با بکارگیری شبکه عصبی پیش‌خور با الگوریتم پس انتشار، اقدام به تحلیل تراکنش‌های مشکوک بر اساس الگوهای تقلب می‌پردازد و تقلب تراکنش را در نهایت تایید یا رد می‌نماید. [۴]

شبکه عصبی کوهنن (نگاهداشت خود سازمانده SOM) [۵]

نگاشت‌های خود سازمانده نوع دیگری از شبکه‌های عصبی رقابتی بدون ناظر است. نگاشت‌های خود سازمانده (SOM) مدلی غیر نظارتی بر شبکه‌های عصبی است که توسط کوهنن رشد و بسط یافت، شبکه کوهنن نیز مانند سایر شبکه‌های عصبی دارای ارتباطات کامل و بصورت پیش خور می‌باشد و پیشخور بودن آن به این معنی که این‌گونه از شبکه ها اجازه ایجاد حلقه را ندارد. همچنین ارتباط کامل شبکه نیز به این معنی که بین هر یک از عصب‌های (گره یا نرون) هر لایه و هریک از عصب های لایه بعد یک ارتباط وجود دارد به هر ارتباطی بین دو لایه عددی، وزنی که مقدار آن بین صفر و یک می‌باشد، بصورت تصادفی تخصیص داده می‌شود، که در واقع تعدیل این وزنها همان مکانیسم یادگیری شبکه های عصبی می‌باشد. اما علی رغم سایر انواع شبکه های عصبی این نوع شبکه دارای لایه پنهان نمی‌باشد و داده ها بصورت مستقیم از لایه ورودی به لایه خروجی ارائه می‌شود. لایه خروجی معمولا به شکل یک شبکه در ابعاد یک یا دو می‌باشد، که معمولا بصورت مستطیل شکل بوده و گاها شش گوش نیز می باشند. شبکه کوهنن می‌تواند برای به تصویر کشیدن داده های ابعادی بال به کار گرفته شود. این نگاشت سطوحی را که در فضای ورودی به هم نزدیک‌اند به فضایی که در آن داده ها در فضای خروجی نزدیکند، تصویر می‌کند. شکل(۱) نشان می دهد که چگونه یک بردار ورودی هفت بعدی به یک فضای اصلاح شده دو بعدی تصویر شده است.



شکل ۱: کاهش ابعادی در نگاشت خود سازمانده

اگرچه شبکه کوهنن یکی از روش‌های کاهش بعد نیز به حساب می‌آید اما کاربرد اصلی این شبکه در تحلیل خوشه(خوشه بندی) می باشد. بطوری‌که این شبکه قابلیت تفکیک داده های ورودی را به گروه هایی جدا از هم، که بر اساس شباهت‌ها و الگوهای پنهان بین این داده ها تقسیم‌بندی می‌شوند را دارا می‌باشد.

بردار ورودی کوهنن را بصورت مجموعه ای از مقادیر M فیلد برای N امین رکورد بصورت زیر در نظر می‌گیریم ...,xn2,xn1 و $An=xnm$ که مطابق با آن M وزن برای لایه خارجی Jم نیز بصورت بردار $wj = w, w2,.....wmj$ در نظر گرفته می شود.

الگوریتم شبکه کوهنن طی مراحل زیر صورت می پذیرد.

بایستی برای هر بردار ورودی X مراحل زیر را انجام می دهیم:

رقابت: برای هر گره خروجی J ام مقداری برای D(Wj , Xn) به عنوان تابع ارزیابی محاسبه می کنیم برای مثال با فرض فاصله اقلیدسی D بصورت زیر بدست می آید.

$$D(wj,Xn)=\sqrt{(\sum iWij-Xni)}$$

گره ای بعنوان گره برنده معرفی می گردد که مینیم فاصله را از بین گره ها ورودی از گره J ام داشته باشد.

همکاری: در این مرحله اقدام به شناسایی همه خروجی گره J ام که در همسایگی تعریف شده از گره j بوسیله شاخص اندازه همسایگی که آن را با حرف R نشان می دهند، اقدام می نمائیم و سپس برای همه این گره ها عملیات بالا را تکرار می کنیم تا تعدیلی دراوزان به وجود آید.

انطباق: تنظیم نرخ یادگیری و اندازه همسایگی در صورت نیاز

هنگامی که شرایط خاتمه مهیا گردد **توقف** می نمائیم.

آکادمی داده
dataacademy.ir

جمع آوری اولیه داده

داده های اولیه جهت بررسی و پیاده سازی فرایند داده کاوی مربوط به تراکنشهای صورت گرفته از طریق کارتهای صادره بانک ملت می باشد، بعد از بررسی های انجام شده در بانک مورد نظر، داده های مربوط به کارت اکسس کارت این بانک مرود انتخاب قرار گرفت، دلیل انتخاب این کارت بیان این موضوع می باشد، که به هر دارنده صاحب کارت یک شماره حساب بانکی نیز تخصیص داده شده و همچنین این کارتها قابلیت انجام تراکنش با مبالغ بالا و از طریق مختلف (اینترنت، خودپرداز و پوز) را دارا می باشد، برای مثال بعضی از کارتهای بانکی سیستم شتاب دارای این ویژگی ها نبوده و سقف تراکنش‌های انجام شده روزانه برای آنها حداکثر مبلغ ۱۲۰۰۰۰ تومان می باشد. به علت حجم بالای کلیه داده های تراکنشهای ذخیره شده در پایگاه اطلاعاتی بانک و همچنین تصمیم مدیران ارشد بانک در خصوص ارائه این سرویس ویژه برای افرادی خاص که دارای درجه اهمیت بالاتری نسبت به سایر دارندگان کارتهای بانکی می باشند، داده های مربوط به ۱۰۰ نفر اول از لحاظ حجم مبادلات ریالی صورت گرفته از طریق کارت‌های آنها در ایران، جمع‌آوری گردیده شد.

توصیف داده

بعد از بارگذاری داده ها در نرم افزار بانک اطلاعاتی تعداد تراکنشهای ثبت گردیده (تعداد رکوردها) برای ۱۰۰ مشتری انتخاب شده حدود ۷۴۰۰۰ رکورد مشخص گردید همچنین فیلدهای اطلاعاتی موجود در پایگاه داده ابتدایی، بصورت فیلدهای نام برده شده درجدول شماره ۱ مشخص گردیده شد.

جدول ۱: مشخصات پایگاه داده تراکنش ها

نام ستون (Column Name)	نوع داده (Data Type)	طول رشته (String length)	تشریح فیلد (Description)
PAN	Varchar	۲۰	شماره کارت
TRANSNO	Varchar	۲۰	شماره تراکنش
TRANSDATE	Numeric	Integer	تاریخ تراکنش
DEVICENO	Numeric	Integer	کد دستگاه
DEVICETYPE	TEXT	۵	نوع تراکنش
AMOUNT	Numeric	Long Integer	مقدار تراکنش
RESULTTRANS	TEXT	۵	نتیجه تراکنش

اکتشاف داده

بعد از بارگزاری داده ها در نرم افزار داده کاوی و بررسی و تحلیل داده ها در قالب گرافهای توزیع مختلف بر اساس فیلدهای اطلاعاتی، متوجه وجود ارتباط معنی داری بین مبلغ تراکنش و دستگاه مورد استفاده و نوع تراکنش انجام شده، شدیم، به اینصورت که مبالغی که به تعداد بالا تکرار شده بودند و یا دارای مقادیر بالایی بودند، بیشتر از طریق پوز و اینترنت انجام گردیده شده اند.

انتخاب داده

با توجه به اینکه فرض اولیه مساله در نظر گرفتن داده های یک فرد خاص به عنوان رفتاری قانونی بدون داشتن آگاهی قبلی نسبت به احتمال وقوع تخلف در تراکنشهای او می باشد، در این مرحله اقدام به انتخاب یک نفر با رفتاری نرمال تر و با توزیع متنوع تر نسبت به بقیه از بین صد نفر، شد، به این معنی که حدالمکان در تراکنش های آن قرد خاص، سه نوع تراکنش های اینترنت و پوز و دستگاه های خود پرداز و همچنین تراکنش رد و قبول نیز به نسبتهای نزدیک به هم، موجود باشد. در نهایت فردی دارای حدوداً ۳۲۰ تراکنش انجام شده در بازه زمانی دو سال انتخاب شد.

پاکسازی داده ها

بعد از بررسی های صورت گرفته بر روی پایگاه اطلاعاتی موجود، به دلیل فرایند ذخیره سازی کاملاً خودکار داده ها در بانک، رکوردهایی با مقادیر گمشده ملاحظه نگردید. اما دو رکورد به عنوان نویز با مقادیر اشتباه برای داده ها مشخص گردید (کد دستگاه ۱۰۰ اشتباهاً در یک تراکنش ۰۰۰۱۰۰ و کد دستگاه ۱۱۷۱۱ اشتباهاً در دو تراکنش ۰۰۰۱۱۷۱۱ ثبت شده بود) که بصورت دستی اقدام به اصلاح و یا به تعبیر دیگر هموارسازی داده ها نمودیم. از طرفی ۷ رکورد نیز به عنوان داده های پرت شناسایی شد که به جهت احتمال تشخیص آنها توسط مدل مورد نظر، جهت کشف رفتار مشکوک، اقدامی برای خروج آنها از داده ها صورت نپذیرفت.

ساخت داده

در این مرحله خصوص عدم قابلیت استفاده الگوریتم خوشه بندی شبکه عصبی کوهنن از فیلدهای اطلاعاتی کد دستگاه به دلیل حروفی بودن نوع قابل بندی و تنوع زیاد مقادیر این متغیر ورودی و همچنین فیلد اسطلاعاتی تاریخ انجام تراکنش به دلیل تنوع زیاد آن و قالبندی عددی برای آن در پایگاه داده موجود، با استفاده از نرم افزار بانک

آکادمی داده
dataacademy.ir

مقادیر شمارش شده از تکرار هر دستگاه مشخص در کل رکوردها به جای عدد کد دستگاه در همه رکوردهای پایگاه داده، جایگزین شده و فیلد اطلاعاتی جدیدی با نام Count-Deviceno ایجاد گردید.

هر تراکنش بر اساس تاریخ انجام تراکنش در یکی از شش گروه تاریخی تراکنشهای پنج روزه اول تا پنج روزه ششم ماه طبقه بندی شده و مقادیر مربوط به آن در فیلد اطلاعاتی جدیدی با نام fiveday قرار گرفت.

تقسیم بندی داده ها

بعد از انتخاب و آماده سازی داده ها، داده ها بصورت تصادفی به سه مجموعه داده مجزا با نام مجموعه داده های آموزشی، آزمون و اعتبار سنجی تقسیم بندی شدند. کاربرد این مجموعه داده ها را می توان بترتیب جهت ساخت اولیه یک مجموعه مدل، مشخص نمودن عملکرد مدل ساخته شده بروی داده ها دیده نشده و انتخاب بهترین و کاراترین مدل از میان کلیه مدلها، دانست. میزان درصد حجم تراکنشهای مجموعه داده آموزشی ۷۵٪، آزمایشی ۲۰٪ و اعتبار سنجی ۵٪ می باشد.

آماده سازی داده های مشکوک

چون از ابتدای تحقیق پایگاه داده مربوط به تراکنشهای کارتهای بانکی برای هر فرد بعنوان رفتار قانونی آن فرد، فرض شده است، بنابراین به دلیل ارزیابی مدل در خصوص شناسایی رفتار مشکوک در این مرحله۲۰ رکورد داده جدید با رفتاری مشکوک در پایگاه داده قبلی اضافه گردیده شد. ساز و کار تهیه تراکنشهای مشکوک بصورت دستی و با اعمال تغییراتی بصورت معنی وار بر روی داده ها بود است

تعیین متغیرهای ورودی

در این مرحله فیلدهای اطلاعاتی نتیجه تراکنش(X1)، مقدار تراکنش(X2)، نوع تراکنش (X3)، شمارش کد دستگاه (بجای کد دستگاه) (X4) و پنج روزه تاریخ انجام تراکنش (به جای روز انجام تراکنش) (X5) به عنوان پنج متغیر ورودی جهت ساخت مدل‌های خوشه‌بندی مورد انتخاب قرار گرفت.

مدل سازی و ارزیابی جهت انتخاب مدل بهینه

ساخت مدل‌های خوشه بندی

در این مرحله با استفاده نرم افزار داده کاوی اقدام به طراحی مدلهای خوشه بندی مختلفی با تکنیک شبکه عصبی کوهنن شد، ۳ مدل با استفاده از همان ۵ متغیر ورودی تشریح شده (X1 , X2,…,X5) و مجموعه داده های ورودی آموزشی و ارائه تنظیمات مختلف بر روی این مدلها مطابق با جدول ۲ طراحی شدند.

جدول ۲: مشخصات مدل‌های تولید شده شبکه عصبی کوهنن

شماره مدل	تعداد خوشه	لایه ورودی(نرون)	لایه خروجی(نرون)	تنظیمات(طول×عرض)	نوع مدل
مدل۱	هشت	۱۲	۹	۳×۳	خبره
مدل۲	نه	۱۲	۹	۳×۳	ساده
مدل۳	یازده	۱۲	۱۲	۳×۴	ساده

تولید یک طرح آزمون جهت ارزیابی مدلها و انتخاب مدل بهینه

جهت ارزیابی نتایج پروژه های داده کاوی، روشهای مشخص و متنوعی جهت ارزیابی مدلهای نظارتی به ویژه مدلهای طبقه بندی، موجود می باشد، این موضوع برای مدلهای غیر نظارتی صدق نمی کند، در مقابل یکی از مهمترین روشهای ارزیابی مدلهای غیر نظارتی، تبدیل اینگونه مسائل به یک مدل نظارتی و استفاده از روشهای ارزیابی مرتبط با اینگونه از مدلها می باشد.[۸ و۹ و۱۰] در تحقیق حاضر نیز از این مفهوم، جهت تبدیل هر مدل خوشه یندی به یک مدل طبقه بندی و ارزیابی از طریق روشهای ارزیابی مدلهای طبقه بندی، استفاده گردیده شد. در همین خصوص گامهای زیر جهت ارائه یک الگوریتم ارزیابی، پیشنهاد می گردد:

بعد از خوشه بندی داده ها، هر خوشه بدست آمده در این مدل را بعنوان یک کلاس در نظر گرفته و در صورت عدم امتخاب خودکار اسم برای آن، هر کلاس را با اسامی مانند...، Cluster2، Cluster1 نامگذاری کرده و مجموعه تراکنشهای عضو آن خوشه را متعلق به کلاس نام گذاری شده متناظر آن نیز می دانیم.

در این مرحله توسط همان داده های آموزشی ابتدایی اقدام به آموزش متناظر مدل مجموعه قوانین C5 برای هر یک از مدلهای خوشه بندی می گردد. لازم به ذکر است فیلدهای اطلاعاتی ورودی مدلهای خوشه بندی (X1,X2,…,X5) به عنوان فیلدهای ورودی مدلهای طبقه بندی و اسامی کلاسهای تخصیص داده شده به عنوان متغیر خروجی در ملهای طبقه بندی می باشد.

در این مرحله با عبور مجموعه داده های آزمون و ارزیابی از مجموعه مدلهای خوشه بندی و طبقه بندی مجموعه قوانین، میزان قابلیت اطمینان مدل‌ها محاسبه می گردد.

با ارزیابی مقادیر شاخص های قابلیت اطمینان مدلها مطابق با جدول ۳، مدل بهینه انتخاب می گردد.

آکادمی داده
dataacademy.ir

جدول(۳): معیارهای ارزیابی مدل‌ها

ردیف	تشریح معیار	شاخص
۱	قابلیت اطمینان کل مدل طبقه بندی برای مجموعه داده های آموزشی	تعداد کل طبق بندی صحیح / تعداد کل داده های ارائه گردیده
۲	قابلیت اطمینان کل مدل طبقه بندی برای مجموعه داده های آزمایشی	تعداد کل طبق بندی صحیح / تعداد کل داده های ارائه گردیده
۳	قابلیت اطمیان کل مدل طبقه بندی برای مجموعه داده های ارزیابی	تعداد کل طبق بندی صحیح / تعداد کل داده های ارائه گردیده
۴	بهبود در روند شاخص های قابلیت اطمینان برای مجموعه داده های ارزیابی نسبت به داده های آزمون نسبت به داده های آموزشی	میزان روند رو به رشد

ارزیابی نهایی و انتخاب مدل بهینه

با توجه به نتایج قابلیت اطمینان کل مدلها که در جدول(۴) مشخص می باشد، به دلیل مقدار کم و روند نزولی شاخص قابلیت اطمینان کل، مدل شماره ۱ را به عنوان مدل نهایی جهت کشف رفتار مشکوک انتخاب می کنیم.

جدول (۴): مقایسه قابلیت اطمینان کل ۳ مدل با روش شبکه عصبی کوهنن

شماره مدل	شاخص قابلیت اطمینان با استفاده از داده های آموزشی(درصد)	شاخص قابلیت اطمینان با استفاده از داده های آزمون(درصد)	شاخص قابلیت اطمینان با استفاده از داده های ارزیابی(درصد)
مدل۱	۱۰۰	۱۰۰	۱۰۰
مدل۲	۹۹.۵۶	۹۶.۹۲	۹۶.۳
مدل۳	۹۹.۵۶	۱۰۰	۹۲.۵۹

شناسایی رفتار مشکوک

در مدلهای طراحی شده اعدادی در خروجی مدلهای طبقه بندی مجموعه قوانین C5 به عنوان مقادیر احتمال قابلیت اطمینان تخصیص یک رکورد به کلاس تخصیص داده شده قابل محاسبه و استفاده می باشد، به این صورت فیلد جدیدی که با عبارت $CC شروع می گردد و حاوی این مقادیر می باشد، در پایگاه واده موجود ایجاد می شود حال جهت شناسای رفتار مشکوک از این مقادیر استفاده می کنیم. مینیم مقدار آن اعداد را در داده های آموزشی پیدا کرده و در صورت اینکه مقادیر این شاخص برای داده های جدید(برای مثال داده های آزمایشی) کمتر از آن مقدار مینیمم باشد، آن تراکنش (رکورد) را به عنوان تراکنش مشکوک(رفتار مشکوک) در نظر می گیریم. مدل حاضر ۱۷ رفتار را به عنوان رفتار مشکوک شناسایی کرد که ۱۴ عدد از تراکنشها شناسایی شده از همان ۲۰ تراکنش مشکوک تهیه شده قبلی به روش دستی می باشد، این مقدار خود نیز نشان دهنده احتمال ۷۰ درصدی شناسایی رفتار مشکوک توسط این مدل می باشد.

استخراج قواعد، تحلیل و نتیجه‌گیری

در این مرحله ۱۳ قانون مربوط به مدل طبقه بندی مجموعه قوانین C5 متناظر با شبکه کوهنن با تعداد ۸ خوشه را استخراج نموده‌ایم و با بررسی های بیشتر بر روی مکانیزم تولید هر یک از این قوانین، رفتارهای مشکوک شناسایی شده توسط این مدل و همچنین نکات دیگر بدست آمده در حین تحقیق، بصورت کلی نتایج زیر حاصل گردید.

نتایج فرد گرایانه همچون: اگر شخصی با کارت فرد مورد نظر از تاریخ ششم ماه تا تاریخ ۱۰ ماه تراکنش از طریق پوز انجام دهد این تراکنش مشکوک به نظر می رسد چون معمولاً در مدل رفتاری این فرد هیچگاه چنین تراکنش انجام نشده است، و یا اینکه فرد مذکور همیشه از طریق یک ترمینال خاص از طریق اینترنت تراکنش مالی انجام رفتارهای مشکوک قرار گیرد و موارد دیگر...

آن دسته از خصوصیات متغیر های ورودی که در کل داده های تاریخی، دارای سهم(درصدی) پایینی می باشند، دارای احتمال بالایی برای تخصیص نسبتهای بالا به خود، درکل رفتار مشکوک شناسایی شده هستند. برای مثال تعداد تراکنشهای POS که در کل داده های آموزشی سهم ۳ درصدی دارا بوده است، درکل رفتار مشکوک شناسایی شده توسط مدل، سهمی معادل با ۵۳ درصد را دارا می باشد.

عدم وجود یک فیلد در متغیرهای ایجاد کننده شروط در قوانین، دو معنی می تواند داشته باشد، معنی اول تنوع زیاد آن متغیر ورودی و عدم ایجاد الگو توسط مدل برای آن به دلیل تنوع زیاد می باشد و معنی دوم وجود رفتاری یکسان و الگوی بیش از اندازه شفاف و بدیهی برای آن متغیر می باشد.

آکادمی داده
dataacademy.ir

از تکنیک شبکه عصبی کوهنن می توان جهت غربال‌گری اولیه داده ها در سیستمهای کشف رفتار مشکوک در زمان حقیقی استفاده نمود ه تعبیری می توان ابتدا یک سری از داده ها را توسط این مدل از کل تراکنشهای صورت گرفته به عنوان رفتارهای مشکوک انتخاب نموده و سپس توجه خود را بصورت دقیق تر بر این داده ها جهت کشف رفتار متقبلانه معطوف نمائیم، این موضوع باعث کاهش هزینه ها و افزایش سرعت و امکان کشف رفتار متقلبانه در زمان حقیقی می گردد.

مدل طراحی شده در این تحقیق فرد گرا بوده و نتایج مرتبط با آن نیز به میزان بالایی متکی بر رفتار افراد می باشد و در صورت ارائه داده های افراد دیگر به مدل احتمال وجود نتایج متفاوت وجود دارد، اما می توان اینگونه بیان نمود که مدل طراحی شده فعلی قابلیت تعمیم و استفاده برای مشتریانی با رفتاری شبیه به رفتار فرد مذکور را دارا می باشد.

یکی از نتایج مهم بدست آمده در این تحقیق وجود الگوهای مختلف در فرایند اکتشاف می باشد که این موضوع باعث می گردد که ما با نگاه از طریق مختلف به مساله همواره به دنبال استفاده از تکنیک های متفاوت جهت حل مساله باشیم که نهایتاً این موضوع نیز باعث هم افزایی و کشف دانش بیشتر جهت حل مساله می گردد.

پیشنهاد برای تحقیقات آتی

کشف رفتار مشکوک با رویکرد غیر نظارتی و استفاده از تکنیک های دیگر خوشه بندی مانند Two-K-Means-Step و ...

توسعه مکانیزم ارزیابی مدل‌ای غیر نظارتی و نظارتی با استفاده از شاخص‌ای گسترده‌تر و استفاده از تکنیک‌های تصمیم‌گیری در ارزیابی مدل‌ها جهت انتخاب مدل بهینه کشف رفتار مشکوک کارت‌های بانکی.

ترکیب مدلهای غیر نظارتی و شبکه عصبی پیش خور جهت ارزیابی و انتخاب مجموعه مدل بهینه جهت کشف رفتار مشکوک در تراکنش های کارتهای بانکی.

چگونگی استفاده از مدلهای نظارتی و غیر نظارتی بصورت توام در طراحی سیستمهای مکانیزه کشف تقلب در بانکها، جهت ایجاد هم افزایی در این خصوص.

تحقیق در خصوص طراحی سیستم یکپارچه ثبت و ذخیره سازی اطلات تراکنشها در بانکها جهت تقسیم بندی و برچسب زنی داده های تراکنشها در دو دسته تراکنشهای قانونی و متقلبانه.

منابع

[1] Bolton,R.& Hand, D.Statistical Fraud Detection: A Review(With Discussion). Statistical Science 17(3)
[2] Phua, C. , And Lee, V ., et al. A comprehensive survey of data mining-based fraud detecton reseach,2005.
[3] Zaslavshy,v.,& Strizhak, A .Credit Card fraud detection using self-organizing maps. Information and Security. 18,48-63,2006
[4] Quah, J. T. S., and Sriganesh, M. "Real-time credit card fraud detection using computational intelligence" Expert Systems with Applications35(4):1721-1732,2008
[5] Larose,Daniel T., "Discovering Knowledge In Data Mining", John Wiley & Sons, Inc, Hoboken ,2005
[6] Chapman, P., Clinton, J. ,Kerber, R., Khabaza Th., Rienart, Th., Shearer.C., & Writh R., "CRISP_DM Stepby step Sata Mining Guide"
[7] Gupta, G. k.(2006) Introduction to Data Mining with Case Studies, PHI Learning Pvt. Ltd.
[8] Williams,G. , Huang, Z.(1997) Mining the Knowledhe Mine: The Hot Spots Methodology for Mining Largi Real World Databases. Proc. Of the 10th Australian Joint Conference On Artificial Intelligence
[9] Roiger, R . J. and M. W Geatz(2003) Data Mining: A tutorial-based Primer,Addison Wesley Boston.
[10] Kantardezic, m.,(2003), " Data Mining Concepts, Models,Mothosd and algorithms" John wiley & Sons. Edited by jj, Hoboken.

آکادمی داده
dataacademy.ir