هفتمین همایش سالانه بانکداری الکترونیک و نظام‌های پرداخت
نوآوری، بازیگران جدید و کارآیی در کسب وکار مالی
تهران، مرکز همایش‌های بین‌المللی برج میلاد - ۲و۳ بهمن ۱۳۹۶
7th Annual Conference on Electronic Banking and Payment Systems

ارائه مدلی ترکیبی مبتنی بر خوشه‌بندی و قواعد انجمنی برای شناسایی رفتارهای متقلبانه در تراکنش‌های بانکی

عبدالله عشقی[۱]،مهرداد کارگری[۲]، مصطفی جاویده[۳]، حامد میرشک[۴]، علی محمد نادری[۵]

چکیده:

مسأله تقلب در تراکنش‌های بانکی یکی از مشکلات اساسی در عصر الکترونیکی شدن تراکنش‌های بانکی است. در این مقاله یک مدل ترکیبی نیمه بانظارت با استفاده از الگوریتم‌های خوشه‌بندی و تحلیل انجمنی برای کشف تقلب و رفتارهای مشکوک در تراکنش‌های کارتی بانکی ارائه شده است. مبنای تحلیل، تراکنش‌های نرمال و غیرمتقلبانه مشتریان بوده است. با استفاده از تحلیل انجمنی الگوهای پرتکرار در رفتارهای مشتریان بانکی استخراج شده است. از این الگوها به عنوان قواعد نرمال استفاده شده است که هر تراکنش باید حداقل با یکی از این الگوها مطابقت داشته باشد. در بخش تحلیل رفتار از الگوریتم خوشه‌بندی فازی برای استخراج خوشه‌های رفتاری نرمال هر مشتری استفاده شده است. در صورتی که یک تراکنش ورودی انحراف بالایی از مدل رفتاری نرمال مشتری داشته باشد، در هیچکدام از خوشه‌ها قرار نمی‌گیرد و به عنوان تراکنش پرریسک شناخته می‌شود. نتیجه نهایی از ترکیب نتایج دو بخش تحلیل قواعد و تحلیل روند با استفاده از روش bagging به دست آمده است. نتایج نشان داده‌اند که مدل ترکیبی ارائه شده دقت و صحت بیشتری در کشف موارد مشکوک و متقلبانه داشته است.

کلمات کلیدی: کشف تقلب، خوشه‌بندی، قواعد انجمنی ، kmeans، Apriori

مقدمه

میزان استفاده از خدمات بانکداری الکترونیکی در ۳ سال گذشته رشد قابل توجهی داشته است [1] و به موازات آن آمار تقلب‌ها و کلاه‌برداری‌های صورت گرفته از این مجرا نیز رشد نگران‌کننده‌ای داشته است، به طوریکه حدود ۶۰ درصد از تقلبات انجام شده از طریق کانال‌های الکترونیکی مانند موبایل بانک و اینترنت بانک صورت می‌گیرد [2].

تقلب به عنوان پدیده‌ای نامطلوب، بویژه در صنعت بانک‌داری تأثیرات مخربی بر این صنعت دارد. آمارها نشان می‌دهند که علاوه بر آنکه درصد قابل توجهی از سود در تجارت الکترونیک به دلیل تقلب هدر می‌رود، تقلب‌ها حتی عاملی برای رویگردانی مشتریان در استفاده از سرویس‌های الکترونیکی بانکی به حساب می‌آیند [3]. تلاش‌های فراوانی در صنعت بانک‌داری برای مقابله با پدیده تقلب صورت گرفته است، اما آنچه کشف و پیشگیری کامل از تقلب را تا کنون امکان‌پذیر نساخته‌است، تغییر مداوم رفتار مشتریان بانکی و بالطبع تغییرات رفتاری متقلبان است که در نهایت مدل‌سازی کامل رفتارها ناممکن ساخته است. تغییر مداوم رفتارها باعث می‌شود تا بکارگیری روش‌های مبتنی بر قاعده برای کشف تقلب به تنهایی کافی نباشند [4]، علاوه بر این روش‌های یادگیری ماشین نیز در صورتی که قابل تطبیق با استراتژی‌های جدید نباشند و به صورت ایستا باشند

[۱] - دانشجوی دکتری مهندسی فناوری اطلاعات، دانشگاه تربیت مدرس -دانشکده مهندسی صنایع و سیستم‌ها a.eshghi@modares.ac.ir
[۲] - استادیار دانشکده مهندسی صنایع و سیستم‌ها- دانشگاه تربیت مدرس- گروه مهندسی فناوری اطلاعات m_kargari@modares.ac.ir
[۳] - کارشناسی ارشد مهندسی نرم افزار، شرکت تأمین خدمات سیستم‌های کاربردی کاسپین javide@caspco.ir
[۴] کارشناسی ارشد مهندسی فناوری اطلاعات ، شرکت تأمین خدمات سیستم‌های کاربردی کاسپین mirashk@caspco.ir
[۵] کارشناسی ارشد هوش مصنوعی ، دانشگاه آزاد اسلامی قزوین ° دانشکده برق،رایانه و فناوری اطلاعات am.naderi@qiau.ac.ir

قابلیت چندانی برای کشف تقلب نخواهند داشت [5]. در سال‌های اخیر از روش‌های یادگیری ماشین و تکنیک‌های داده‌کاوی برای کشف تقلب در تراکنش‌های بانکی بسیار استفاده شده است.

از جمله عواملی و مشکلاتی که در کشف تقلب و در فرایند یادگیری الگوریتم‌های کشف تقلب مؤثر هستند، عدم تعادل داده‌های متقلبانه و داده‌های غیر متقلبانه، کنترل هزینه بکارگیری روش‌ها، زمان سریع پاسخ، تعدد ابعاد فضای مورد تحلیل، تغییر مداوم رفتارها، قابلیت یادگیری و عدم وجود داده‌های بانظارت[۶] را می‌توان نام برد [8]–[6]. در این تحقیق روشی نیمه‌بانظارت[۷] برای مدل‌سازی تغییرات رفتاری مشتریان جهت شناسایی رفتارهای ناهنجار و مشکوک به تقلب ارائه شده است.

از جمله موانع اساسی برای گسترش روش‌های کشف تقلب در داده‌های بانکی عدم تبادل دانش در این زمینه است. هرچند که تحقیقات دانشگاهی قابل توجهی در این زمینه صورت گرفته است، اما تعداد کارهایی که ادعای عملیاتی شدن داشته باشند بسیار کم هستند [9]. عدم وجود مجموعه داده‌های بانظارت استاندارد برای ارزیابی و یادگیری مدل‌ها و الگوریتم‌ها یکی از موانعی است که بویژه در ایران بسیار قابل توجه است. این مشکلات اهمیت توجه به روش‌های داده‌کاوی بی‌نظارت در مسأله کشف تقلب را بیشترمشخص می‌سازند. به همین دلیل در این تحقیق به ارائه یک روش نیمه‌بانظارت مبتنی بر خوشه‌بندی و قواعد انجمنی به صورت ترکیبی پرداخته شده است، تا بر اساس داده‌های تاریخی مشتریان و با توجه به روند رفتاری نرمال آنها در تراکنش‌های گذشته، تراکنش‌های مشکوکی که احتمالاً توسط خود مشتری صورت نگرفته است و ممکن است متقلبانه باشند شناسایی شوند.

ادبیات موضوع

از لحاظ نوع داده‌های قابل تحلیل همه روش‌های کشف تقلب را می‌توان در سه دسته اصلی روش‌های بانظارت، روش‌های نیمه‌بانظارت و روش‌های بانظارت قرار داد [10]. به دلیل عدم وجود مجموعه داده‌های بانظارت و همچنین بعضی از کاستی‌هایی که روش‌های بانظارت برای کشف تقلب دارند [10]، در سال‌های اخیر توجه ویژه‌ای به روش‌های بی‌نظارت و نیمه‌بانظارت شده است. در بیشتر روش‌های بی‌نظارت از الگوریتم‌های خوشه‌بندی استفاده می‌شود.

تحقیقات [13]–[11] از روش‌های داده‌کاوی بی‌نظارت و الگوریتم خوشه‌بندی kmeans که از جمله الگوریتم‌های مبتنی بر مرکزیت است، برای کشف رفتارهای غیرنرمال و متقلبانه استفاده کرده‌اند. در تحقیق [14] از روش‌های خوشه‌بندی kmeans و brich برای کشف رفتارهای متقلبانه پول‌شویی استفاده شده است.

در تحقیقات [17]–[15] از روش خوشه‌بندی سلسله‌مراتبی در کشف تقلب استفاده شده است. پانیگراهی و همکارانش در تحقیق خود [18] از روش خوشه‌بندی مبتنی برچگالی DBSCAN به همراه روش ترکیبی قواعد دمپستر-شفر برای کشف تقلب در تراکنش‌های بانکی استفاده کرده‌اند.

در [19] از یک روش خوشه‌بندی با الگوریتم k-means بهبودیافته استفاده شده است. در این روش خوشه‌بندی بر روی تراکنش‌های تاریخی مشتریان صورت گرفته است و تراکنش‌های جدید چنانچه به هیچکدام از خوشه‌های قبلی پیدا شده برای مشتری تعلق نداشته باشند به عنوان تقلب شناسایی می‌شوند. بر اساس گزارش SAS [20] هر سیستم کشف تقلب پیشرفته

---

[۶] Supervised

[۷] Semi-supervised

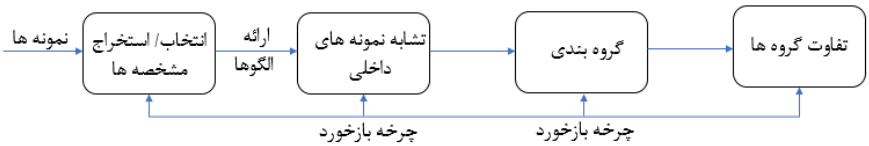
در کنار بخش مبتنی برقواعد لازم است یک بخش تحلیلی نیز داشته باشد که این بخش بر اساس تحلیل‌های بی‌نظارت یا نیمه بانظارت قرار دارد. تقلب‌های جدید و ناشناختانه توسط تحلیل‌های بی‌نظارت یا نیمه‌بانظارت قابل شناسایی هستند.

در [21] یک روش نیمه‌بانظارت با استفاده از داده‌های غیرمتقلبانه کاربران ارائه شده است. در این روش ابتدا با استفاده از روش قواعد انجمنی و الگوریتم Apriori تعدادی قاعده استخراج شده است. این قاعده‌ها بر روی داده‌های غیرمتقلبانه کاربران اعمال شده است، قواعدی که با این داده‌های مطابق بوده‌اند حذف شده‌اند. از قواعد باقی‌مانده برای مانیتور شدن سیستم اصلی استفاده شده است و قواعدی که قادر به تشخیص موارد پرت نبوده‌اند حذف شده‌اند، قواعد باقیمانده با استفاده از ایجاد جهش‌های کوچک در آنها تکرار شده‌اند. از این روش به صورت عملیاتی و برای کشف تقلب‌های داخلی استفاده شده است.

در [22] با استفاده از الگوریتم‌های خوشه‌بندی، پروفایل رفتاری نرمال مشتریان استخراج شده است و از آنها برای یافتن رفتارهای مشکوک مشتریان استفاده شده است. در [23] از یک روش نیمه‌بانظارت شبکه عصبی با یک لایه پنهان و تعداد یکسانی نورن‌های ورودی و خروجی برای تراکنش‌های غیرمتقلبانه هر کارت اعتباری استفاده شده است. در [24] کاربرد روش‌های کشف داده‌های پرت برای کشف تقلب بررسی شده است و روش‌های مبتنی بر دسته‌بندی و خوشه‌بندی و همراه با نحوه کاربرد آنها شرح داده شده است.

خوشه‌بندی

خوشه‌بندی یکی از شاخه های یادگیری بی‌نظارت است و فرآیندی است که در طی آن، نمونه‌ها به دسته‌هایی معنی‌دارکه اعضای آن مشابه یکدیگر هستند تقسیم می‌شوند که به هرکدام از این دسته‌ها خوشه گفته می‌شود. خوشه‌هایی که در اثر این دسته‌بندی به حاصل می‌شوند لازم است تا همگن باشند. در یک خوشه‌بندی خوب، نمونه‌های داخل یک خوشه بسیار به هم شبیه و نمونه‌های خوشه‌های مختلف خیلی با هم متفاوت هستند. این فرایند شامل مراحلی است که در شکل ۱ نشان داده شده است.



شکل ۱- مراحل اصلی در خوشه‌بندی

الگوریتم‌های خوشه‌بندی به دو دسته اصلی الگوریتم‌های بخش‌بندی و الگوریتم‌های سلسله‌مراتبی تقسیم‌بندی می‌شوند [25] که آنها نیز به الگوریتم‌های مبتنی بر مرکزیت، الگوریتم‌های مبتنی بر پیوستگی، الگوریتم‌های مبتنی بر چگالی، الگوریتم‌های مفهومی و الگوریتم‌های مبتنی بر یک تابع هدف تقسیم می‌شوند.

قواعد انجمنی

تحلیل انجمنی (وابستگی)، یکی از الگوریتم‌های داده‌کاوی است که با استفاده از آن می‌توان صفات یا ویژگی‌هایی از یک پدیده یا داده را که با هم می‌آیند مطالعه مطالعه نمود. با استفاده از این روش می‌توان وابستگی بین یک یا چند مشخصه را کشف کرد و به این ترتیب می‌توان روابط بین یک یا چند مشخصه را کمی سازی نمود[26] .

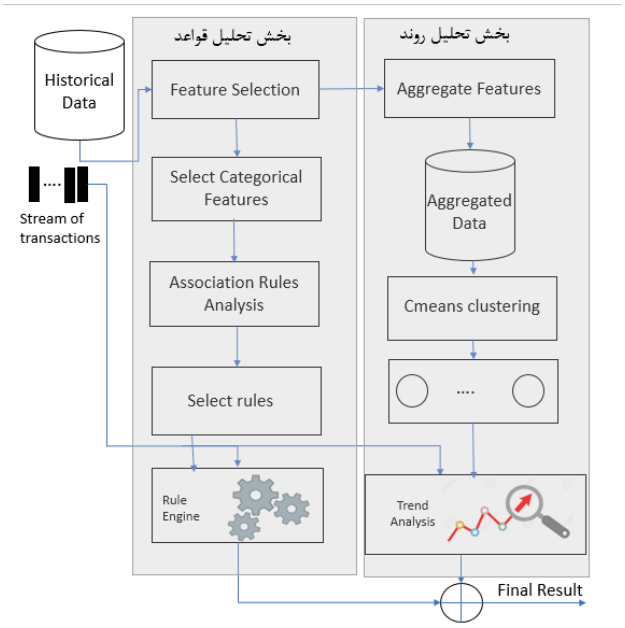
قواعد انجمنی (قواعد وابستگی) به شکل "اگر قسمت مقدم قاعده آنگاه قسمت تالی قاعده" تعریف می‌شوند. به عنوان مثال اگر شخصی مرد/زن از طریق کانال موبایل در ساعات آخر روز در صنف کتاب‌فروشان تراکنشی داشته باشد، آنگاه مقدار این

تراکنش کمتر از x ریال است و این را می‌توان به عنوان یک قاعده تعریف کرد.

از جمله روش‌های قواعد انجمنی می‌توان روش Apriori، روش Eclat و روش FP-growth را نام برد.

روش تحقیق

روش ارائه شده در این تحقیق شامل دو بخش اصلی است که عبارتند از: بخش مبتنی بر قاعده و بخش تحلیل روند. ساختار کلی مراحل کار به صورت شکل ۲ است.



شکل ۲- ساختار کلی مدل پیشنهادی کشف تقلب

هر تراکنش پس از ورود به سیستم، از هر دو بخش عبور کرده و نتیجه هر بخش به صورت مستقل محاسبه شده و در نهایت با استفاده از روش بگینگ[۸] هم ترکیب می‌شوند. خروجی هرکدام از بخش‌ها یکی از دو مورد کم ریسک و پر ریسک است و نتیجه نهایی یکی از سه مقدار کم ریسک، ریسک متوسط و پرریسک خواهد بود. در جدول ۱ نتیجه ترکیب دو بخش مبتنی برقاعده و تحلیل روند و نتیجه نهایی نشان داده شده است.

جدول ۱- ترکیب ریسک ها و نتیجه نهایی با استفاده از روش بگینگ

خروجی بخش تحلیل روند	خروجی بخش مبتنی بر قاعده	خروجی نهایی
کم ریسک	کم ریسک	**کم ریسک**
پر ریسک	کم ریسک	**ریسک متوسط**
کم ریسک	پر ریسک	**ریسک متوسط**
پر ریسک	پر ریسک	**پر ریسک**

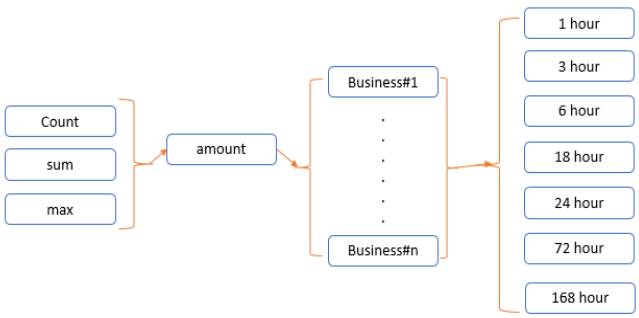
[۸]- bagging

یکی از مراحل مهم در الگوریتم‌های کشف تقلب انتخاب مشخصه‌های درست و تأثیرگذار است [2]. مشخصه‌های قابل استفاده را می‌توان به لحاظ نوع استخراج آنها به دو دسته مشخصه‌های اصلی و مشخصه‌های مشتق شده دسته‌بندی کرد. مشخصه‌های اصلی آنهایی هستند که در همه تراکنش‌ها وجود دارند لیستی از این مشخصه‌های در جدول ۲ نشان داده شده است.

جدول ۲- مشخصه‌های اصلی موجود در تراکنش

	نام مشخصه	توضیحات
۱	Transaction ID	مشخصه تراکنش
۲	Time	زمان و تاریخ انجام تراکنش
۳	Account number	مشخصه شناسایی حساب مشتری
۴	Card number	مشخصه شناسایی کارت
۵	Transaction type	کانال انجام تراکنش مثلاً POs،ATM و ...
۶	Entry mode	تراکنش با حضور کارت یا بدون کارت
۷	Amount	مقدار تراکنش
۸	Merchant code	کد شناسایی پذیرنده
۹	Merchant group	نوع و صنف پذیرنده
۱۰	Gender	جنسیت دارنده کارت
۱۱	Age	سن دارنده کارت
۱۲	Bank	بانک صادرکننده کارت

مشخصه‌های مشتق شده به طور عمومی در متن تراکنش نیستند و نیاز به انجام پردازش برای محاسبه آنها وجود دارد. این مشخصه‌ها خود به دو دسته مشخصه‌های ساده و مشخصه‌های تجمیعی دسته‌بندی می‌شوند. مشخصه‌های تجمیعی نقش بسیار مهمی در تحلیل روند رفتاری و الگوریتم‌های کشف تقلب دارند [27] ,[2]. این مشخصه‌ها عددی هستند و میزان یا تعداد تراکنش‌های انجام شده توسط یک کارت را در دوره‌های زمانی مشخص شده نشان می‌دهند. در این تحقیق مشخصه‌های تجمیعی مقدار، تعداد و بیشینه تراکنش‌هایی که توسط یک کارت در بازه‌های زمانی ۱ ساعت، ۳ ساعت، ۱۲ ساعت، ۲۴ ساعت، ۳ روز، یک هفته و یک ماه توسط کانال‌های مختلف انجام می‌شوند محاسبه شده و در تحلیل مورد استفاده قرار گرفته است. شکل ۳ نحوه استخراج مشخصه‌ها را نشان می‌دهد. بازه‌های زمانی انتخاب شده با توجه به توصیه‌هایی است که در تحقیق‌های [27] ,[2] مطرح شده است. این مشخصه‌ها عددی هستند و در بخش تحلیل روندها مورد استفاده قرار گرفته‌اند. ازجمله مشخصه‌های ساده مشتق شده میانگین، ماکزیمم و انحراف معیار مقدارها را می‌توان نام‌برد.



شکل ۳- نحوه استخراج مشخصه‌ها در بازه‌های زمانی و از طریق کانال‌های مختلف

بخش مبتنی بر قواعد

برای تحلیل این بخش از قواعد انجمنی استفاده شده است. مشخصه‌های مورد استفاده برای این بخش همانطور که در شکل ۲ نیز نشان داده شده است، مشخصه‌های دسته‌ای هستند. اولین مرحله در این بخش آماده سازی داده است و مجموعه داده را به شکلی که قابل استفاده توسط الگوریتم‌های قواعد انجمنی باشد، تبدیل می‌کنیم به صورتی که هر مشخصه به انواع قابل تقسیم آن گسسته‌سازی شده و مقدار آن در هر تراکنش ۰ به معنی عدم وجود آن مشخصه یا ۱ به معنی وجود آن مشخصه خواهد شد. جدول زیر نمونه‌ای از داده آماده سازی شده برای انجام تحلیل‌های مبتنی بر قواعد را نشان می‌دهد.

جدول ۳- نمونه‌ای از داده‌های آماده‌سازی شده برای تحلیل انجمنی

ردیف	نام مشخصه	تراکنش		
۱	زن	۰	۱	۰
۲	مرد	۱	۰	۱
۳	نوجوان	۰	۰	۰
۴	جوان	۰	۰	۱
۵	میانسال	۱	۰	۰
۶	پیر	۰	۱	۰
۷	زمان تراکنش/ از ۰ تا ۷	۰	۱	۰
۸	زمان تراکنش/ از ۷ تا ۱۳	۱	۰	۰
۹	زمان تراکنش/ از ۱۳ تا ۲۱	۰	۰	۱
۱۰	زمان تراکنش/ از ۲۱ تا ۲۴	۰	۰	۰
۱۱	کانال موبایل	۰	۱	۰
۱۲	کانال POS	۱	۰	۰
۱۳	کانال اینترنت	۰	۰	۱
۱۴	کانال ATM	۰	۰	۰
۱۵	نوع پذیرنده ۱	۱	۰	۰
۱۶	....	...	۰	۱
۱۷	نوع پذیرنده ۱۰	۰	۱	۰
۱۸	همان پذیرنده قبلی	۰	۰	۰
۱۹	همان کانال قبلی	۰	۱	۰
۲۰	مقدار تراکنش کم	۰	۰	۰
۲۱	مقدار تراکنش متوسط	۰	۰	۱
۲۲	مقدار تراکنش زیاد	۱	۰	۰
۲۳	مقدار تراکنش خیلی زیاد	۰	۱	۰

پس از آماده شدن داده‌های برای استخراج قواعد، الگوریتم Apriori که یکی از الگوریتم‌های قواعد انجمنی است، بر روی داده‌ها اجرا می‌شود. پس از اجرای این الگوریتم زیر مجموعه‌ای از مشخصه‌هایی که بیشترین تکرار را با هم در مجموعه تراکنش‌ها داشته‌اند یافته می‌شوند. در واقع هدف الگوریتم آپریوری، یافتن وابستگی‌ها بین مجموعه‌های مختلف از داده‌است.

خروجی این الگوریتم، مجموعه‌هایی از قوانین است که چگونگی شمول آیتم‌ها در مجموعه‌های داده را توضیح می‌دهد. نمونه‌ای از قواعد استخراج شده توسط این الگوریتم به صورت زیر است:

"اگر مردی نوجوان در ساعت ۲۲ شب از طریق کانال موبایل تراکنشی و بر روی پذیرنده نوع ۴ تراکنشی انجام دهد که تراکنش قبلی آن نه از آن کانال و نه از همان پذیرنده باشد مقدار تراکنش کم خواهد بود."

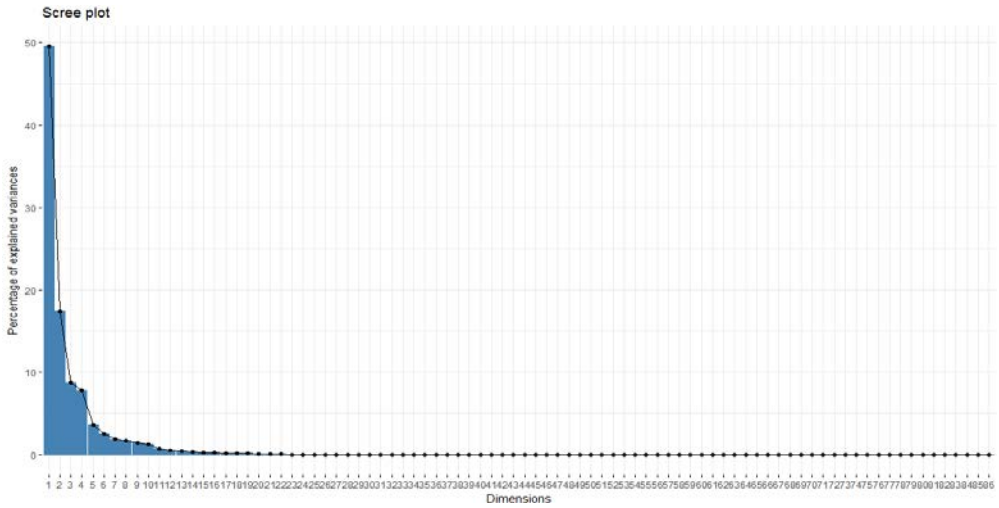
پس از استخراج قواعد، بهینه‌سازی آنها صورت می‌گیرد. برای بهینه سازی قواعد آنها را برروی مجموعه‌داده‌های آزمون که متشکل از تراکنش‌های غیرمتقلبانه هستند اجرا کردیم. قوانینی که منطبق با داده‌های تراکنش‌های مجموعه داده آزمون نبودند را حذف کردیم و مجموعه قواعد باقیمانده را برای تحلیل تراکنش‌های ورودی بکار گرفتیم. هنگامی که یک تراکنش جدید وارد سیستم می‌شود، همه قواعد استخراج شده بر روی آن تست می‌شوند، چنانچه تراکنش با هیچکدام از قوانین موجود منطبق نباشد، آن تراکنش به عنوان پرریسک شناخته می‌شود. وظیفه تطبیق قواعد به عهده موتور قواعد است. نتیجه به دست آمده سپس با نتیجه حاصل از بخش تحلیل روند رفتاری به دست آمده ترکیب شده و نتیجه نهایی حاصل می‌شود.

بخش تحلیل روند

در این بخش از با استفاده از الگوریتم خوشه‌بندی فازی cmeans، خوشه‌بندی بر روی تراکنش‌های نرمال هرکدام از کارت‌ها صورت می‌گیرد و هرکدام از تراکنش‌های نرمال کارت در یک خوشه قرار می‌گیرند. برای اجرای این الگوریتم از مشخصه‌های تجمیعی و عددی تراکنش استفاده می‌کنیم.

با توجه به فرایندی که برای محاسبه متغیرهای تجمیعی در این مقاله تعریف کردیم، تعداد ۸۶ مشخصه جدید استخراج شد که با احتساب مشخصه‌های مانده حساب، مقدار تراکنش و میانگین مانده تعداد مشخصه‌های استفاده شده به ۸۹ مشخصه رسید. برای دقیق‌تر شدن نتایج و پرهیز از نفرین ابعاد، ابتدا یک مرحله کاهش ابعاد بر روی داده‌ها صورت گرفته است. روش‌های مختلفی برای کاهش ابعاد در ادبیات وجود دارد که در [28] مورد بررسی قرار گرفته‌اند. با توجه به ماهیت مشخصه‌های مورد استفاده در این بخش که همه از نوع عددی هستند، در این مقاله از روش کاهش ابعاد PCA[۹] استفاده شده است. در شکل زیر نتیجه اعمال PCA بر روی داده‌های موجود نشان داده شده است.

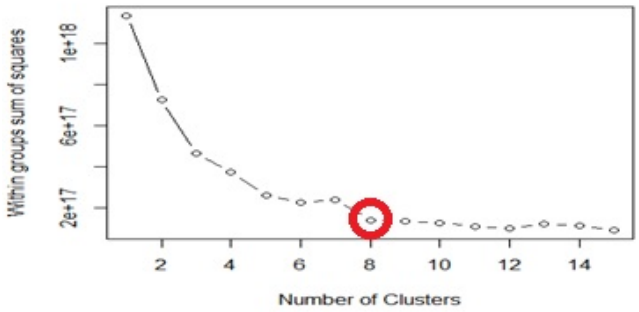
[۹] Principal Component Analysis



شکل ۴- خروجی اعمال PCA بر روی داده‌ها

برای تعیین هر کدام از بخش‌ها به مقدار eigenvalue آنها توجه می‌کنیم، چنانچه بزرگتر از ۱ باشد، نشان‌دهنده این است که PCها واریانس بیشتری در مقایسه با متغیرهای موجود در داده استاندارد سازی شده دارند. از این مقدار به عنوان یک نقطه cutoff برای نگهداشت pc استفاده می‌شود. بر این اساس، تعداد ابعاد قابل استفاده و معنی‌دار برای خوشه‌بندی را ۱۲ انتخاب نمودیم.

در الگوریتم cmeans نمونه‌ها به تعداد c خوشه تقسیم می‌شوند. تعداد c باید از قبل مشخص شده باشد. لازم به ذکر است که برای داده‌های هر کارت یک c متفاوت ممکن است به دست بیاید. برای تعیین اندازه c روش‌های مختلفی وجود دارد که در [28] به تفصیل در مورد آنها بحث شده است. در این تحقیق از روش یافتن زانو در محاسبه SSE[۱۰] استفاده شده است و با رسم نمودار SSE در مقابل تعداد خوشه‌ها و یافتن زانو در شکل می‌توان تعداد خوشه‌های الگوریتم را مشخص کرد. در این روش ابتدا مقدار c را ۲ در نظر می‌گیرد و SSE را محاسبه می‌کند، سپس هر بار به c یک واحد اضافه نموده و مجدداً c را محاسبه می‌کند، چنانچه پس از اضافه شدن تعداد خوشه‌ها تغییر چندانی در SSE ایجاد نشود، مقدار c در آن نقطه را به عنوان تعداد بهینه خوشه‌ها در نظر می‌گیرد. در شکل ۵ روش یافتن تعداد خوشه‌ها برای یکی از کارت‌ها نشان داده شده است که در اینجا تعداد خوشه بهینه برای آن ۸ به دست آمده است.



شکل ۵- محاسبه c برای یکی از کارت‌ها

[۱۰] Sum of Square Error

پس از تعیین تعداد خوشه‌ها، خوشه‌بندی بر روی تراکنش‌ها انجام می‌شود. در جدول ۴ بخشی از خروجی حاصل از اعمال الگوریتم را نشان می‌دهد. همانطور که مشاهده می‌شود هر کدام از تراکنش‌ها با یک درجه‌ای به هر کدام از ۸ خوشه تعلق دارند. تراکنش بیشترین تعلق را به هرکدام از خوشه‌ها داشته باشد، آن خوشه به عنوان خوشه اصلی تراکنش انتخاب می‌شود، مثلاً برای تراکنش شماره ۲۰۲ خوشه اصلی آن خوشه شماره ۳ تشخیص داده می‌شود.

جدول ۴- نتیجه اعمال خوشه‌بندی بر روی تراکنش‌های یک کارت

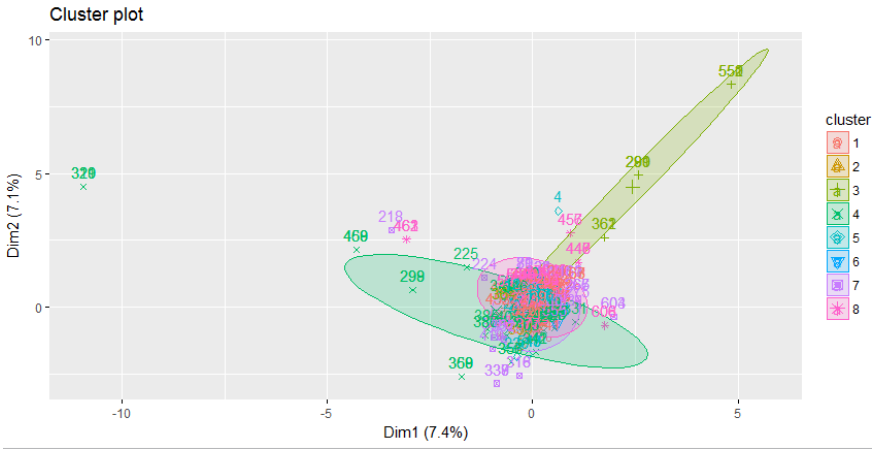
	1	2	3	4	5	6	7	8
202	0.07565578	0.07531216	0.57214645	0.07238493	0.07542144	0.06665178	0.05931105	0.0031164164
551	0.01166859	0.01169947	0.01431993	0.01185935	0.01169018	0.01092711	0.01072896	0.9171064155
286	0.17097992	0.16761352	0.05434246	0.14874751	0.16863930	0.16793644	0.11864531	0.0030955349
615	0.12767577	0.12173979	0.03175306	0.09552415	0.12348869	0.30507318	0.19339987	0.0013454877
654	0.07867602	0.07445176	0.02253146	0.05685035	0.07568711	0.43822453	0.25260020	0.0009785617
32	0.06750322	0.06525850	0.03552227	0.05511506	0.06592094	0.22636271	0.48238600	0.0019312987

پس از آنکه یک تراکنش جدید از یک کارت وارد سیستم شد، میزان درجه تعلق تراکنش به هرکدام از خوشه‌ها محاسبه می‌شود، چنانچه درجه تعلق تراکنش به هیچکدام از خوشه‌های ازقبل محاسبه شده برای آن کارت بزرگتر یا مساوی یک مقدار آستانه تعریف شده توسط خبره سیستم نباشد، آن تراکنش به عنوان تراکنش پرریسک یا پرت شناسایی می‌شود. نتیجه این بخش با نتیجه بخش تحلیل قواعد ترکیب شده و نتیجه نهایی به دست می‌آید.

نتایج و ارزیابی

در این تحقیق از داده‌های تراکنش‌های کارتی یکی از بانک‌های خصوصی ایران استفاده شده است که داده‌های متعلق به ۱۰۰۰ کارت در طول دوره یک ساله از اسفند ۱۳۹۴ تا اسفند ۱۳۹۵ مورد بررسی و تحلیل قرار گرفته است. مجموعه داده متشکل از حدود یک میلیون تراکنش است. برای آزمودن نتایج از یک مجموعه داده برچسب دار شده که توسط خبرگان بانکی به عنوان مجموعه داده تراکنش‌های متقلبانه انتخاب شده‌اند، استفاده شده است. در این مجموعه داده آزمون ، ۷۸۰ تراکنش متقلبانه برای ۱۵ شماره کارت توسط خبرگان بازرسی بانکی شبیه‌سازی شده است.

در شکل ۶ نتیجه اعمال الگوریتم فازی cmeans بروی تراکنش‌های یکی از کارت‌های مجموعه داده نشان داده شده است. همانطور که از شکل ۵ مشاهده می‌شود، تعداد خوشه‌های بهینه برای داده‌ها، ۸ به دست آمده است که در شکل ۶ میزان تعلق هر کدام از تراکنش‌ها به هرکدام از خوشه‌ها مشخص شده است.



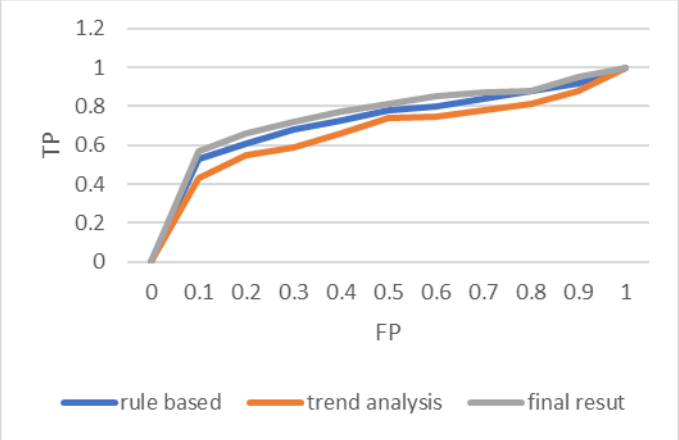
شکل ۶- نتیجه اعمال الگوریتم فازی cmeans بر روی داده‌های تراکنشی کارت

برای ارزیابی مدل استفاده شده، معیار نرخ کشف تقلب و معیار نرخ هشدارهای اشتباه مطابق جدول ۵ مورد بررسی قرار گرفته‌اند. سیستم کشف تقلبی که بیشترین میزان کشف تقلب (TP) را با کمترین میزان هشدارهای اشتباه (FP,FN) داشته باشد مطلوب‌تر است.

جدول ۵- ماتریس پیش‌بینی

		واقعی	
پیشبینی شده		T	F
	T	TP	FP
	F	FN	TN

نتیجه مقایسه هرکدام از بخش‌ها به تنهایی و ترکیب هر دو بخش با استفاده از منحنی ROC در شکل ۷ نشان داده شده است.



شکل ۷- نمودار **ROC** روش های مبتنی برقاعده، تحلیل روند و ترکیبی

همانطور که از نمودار ROC شکل ۷ مشاهده می‌شود، در بخش تحلیل قواعد نسبت به بخش تحلیل روندها که در آن از خوشه بندی استفاده شده است، نتایج بهتری حاصل می‌شود، حال آنکه نتیجه نهایی که از تجمیع نتایج بخش تحلیل روند و بخش مبتنی بر قاعده به دست آمده است نتایج بهتری در مقایسه با هر کدام از روش‌ها به تنهایی دارد به طور نمونه با نرخ هشدارهای اشتباه ۰,۱ (FP=0.1) حدود ۶۰ درصد از موارد تقلب کشف می‌شود این در حالی است که با همین نرخ هشدار اشتباه نتیجه هرکدام از روش‌های مبتنی بر قاعده و تحلیل روند از ۶۰ درصد کمتر است.

با توجه به اینکه کشف تقلب فرایند پیچیده‌ای است و تعداد تراکنش‌های متقلبانه بسیار کمتر از تعداد تراکنش‌های غیرمتقلبانه است، استفاده از تنها یک روش یا الگوریتم قادر به کشف همه موارد تقلب نخواهد بود و لازم است که از روش‌های ترکیبی در سیستم‌های کشف تقلب استفاده شود در این صورت استفاده از روش‌های ترکیب نتایج و گواه‌ها یکی از موضوعاتی است که در تحقیقات آینده کشف تقلب لازم است بیشتر مورد توجه محققان قرار گیرد.

منابع

[1] T. B. Joewono, B. A. Effendi, H. S. A. Gultom, and R. P. Rajagukguk, بInfluence of Personal Banking Behaviour on the Usage of the Electronic Card for Toll Road Payment,پ *Transp. Res. Procedia*, vol. 25, pp. 4454ت4471, Jan. 2017.

[2] A. Correa Bahnsen, D. Aouada, A. Stojanovic, and B. Ottersten, بFeature engineering strategies for credit card fraud detection,پ*Expert Syst. Appl.*, vol. 51, pp. 134ت142, Jun. 2016.

[3] H. Hoehle, E. Scornavacca, and S. Huff, بThree decades of research on consumer adoption and utilization of electronic banking channels: A literature analysis,پ*Decis. Support Syst.*, vol. 54, no. 1, pp. 122ت132, Dec. 2012.

[4] V. Van Vlasselaer *et al.*, بAPATE: A novel approach for automated credit card transaction fraud detection using network-based extensions,پ*Decis. Support Syst.*, vol. 75, pp. 38ت48, Jul. 2015.

[5] A. Dal Pozzolo, O. Caelen, Y.-A. Le Borgne, S. Waterschoot, and G. Bontempi, بLearned lessons in credit card fraud detection from a practitioner perspective,پ*Expert Syst. Appl.*, vol. 41, no. 10, pp. 4915ت4928, Aug. 2014.

[6] M. F. A. Gadi, X. Wang, and A. P. do Lago, بCredit Card Fraud Detection with Artificial Immune System,پin *Artificial Immune Systems*, 2008, pp. 119ت131.

[7] R. J. Bolton, D. J. Hand, and D. J. H, بUnsupervised Profiling Methods for Fraud Detection,پin *Proc. Credit Scoring and Credit Control VII*, 2001, pp. 5ت7.

[8] D. J. Hand, C. Whitrow, N. M. Adams, P. Juszczak, and D. Weston, بPerformance criteria for plastic card fraud detection tools,پ*J. Oper. Res. Soc.*, vol. 59, no. 7, pp. 956ت962, Jul. 2008.

[9] N. Carneiro, G. Figueira, and M. Costa, بA data mining based system for credit-card fraud detection in e-tail,پ*Decis. Support Syst.*, vol. 95, pp. 91ت101, Mar. 2017.

[10] S. Wang, بA Comprehensive Survey of Data Mining-Based Accounting-Fraud Detection Research,پ in *2010 International Conference on Intelligent Computation Technology and Automation*, 2010, vol. 1, pp. 50ت53.

[11] N. A. L. Khac and M. T. Kechadi, بApplication of Data Mining for Anti-money Laundering Detection: A Case Study,پin *2010 IEEE International Conference on Data Mining Workshops*, 2010, pp. 577ت584.

[12] J. Wu, H. Xiong, and J. Chen, بCOG: local decomposition for rare class analysis,پ *Data Min. Knowl. Discov.*, vol. 20, no. 2, pp. 191ت220, Mar. 2010.

[13] R. Liu, X. l Qian, S. Mao, and S. z Zhu, بResearch on anti-money laundering based on core decision tree algorithm,پin *2011 Chinese Control and Decision Conference (CCDC)*, 2011, pp. 4322ت4325.

[14] W. H. Chang and J. S. Chang, بUsing clustering techniques to analyze fraudulent behavior changes in online auctions,پin *2010 International Conference on Networking and Information Technology*, 2010, pp. 34ت38.

[15] L. Torgo and C. Soares, بResource-bounded Outlier Detection Using Clustering Methods,پin *Proceedings of the 2010 Conference on Data Mining for Business Applications*, Amsterdam, The Netherlands, The Netherlands, 2010, pp. 84ت98.

[16] L. Torgo and E. Lopes, بUtility-based Fraud Detection,پ in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*, Barcelona, Catalonia, Spain, 2011, pp. 1517ت1522.

[17] F. H. Glancy and S. B. Yadav, بA computational model for financial reporting fraud detection,پ*Decis. Support Syst.*, vol. 50, no. 3, pp. 595ت601, Feb. 2011.

[18] S. Panigrahi, A. Kundu, S. Sural, and A. K. Majumdar, بCredit card fraud detection: A fusion approach using Dempsterتشafer theory and Bayesian learning,پ*Inf. Fusion*, vol. 10, no. 4, pp. 354ت363, Oct. 2009.

[19] M. Singh and S. Raheja, بCredit Card Fraud Detection by Improving K-Means,پvol. 2, no. 5, 2014.

[20] بSAS Fraud management Report,پ2015.

[21] J. Kim, A. Ong, and R. E. Overill, بDesign of an artificial immune system as a novel anomaly detector for combating financial fraud in the retail sector,پin *The 2003 Congress on Evolutionary Computation, 2003. CEC '03*, 2003, vol. 1, p. 405ت412 Vol.1.

[22] U. Murad and G. Pinkas, بUnsupervised Profiling for Identifying Superimposed Fraud,پin *Principles of Data Mining and Knowledge Discovery*, 1999, pp. 251ت261.

[23] E. Aleskerov, B. Freisleben, and B. Rao, بCARDWATCH: a neural network based database mining system for credit card fraud detection,پin *Proceedings of the IEEE/IAFE 1997 Computational Intelligence for Financial Engineering (CIFEr)*, 1997, pp. 220ت226.

[24] H. Issa and M. A. Vasarhelyi, بApplication of Anomaly Detection Techniques to Identify Fraudulent Refunds,پ Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 1910468, Aug. 2011.

[25] A. Sorin, بSurvey of Clustering based Financial Fraud Detection Research,پ *Informatica Economică*, vol. 16, 2012.

[26] D. T. Larose and C. D. Larose, بAssociation Rules,پin *Discovering Knowledge in Data*, John Wiley & Sons, Inc., 2014, pp. 247ت265.

[27] S. Jha, M. Guillen, and J. Christopher Westland, بEmploying transaction aggregation strategy to detect credit card fraud,پ*Expert Syst. Appl.*, vol. 39, no. 16, pp. 12650ت12657, Nov. 2012.

[28] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques, Third Edition*, 3 edition. Haryana, India; Burlington, MA: Morgan Kaufmann, 2011.