کشف کلاهبرداری بانکی با استفاده از ماشین بردار پشتیبانی

استاد محترم :

جناب اقای امیر شکری

دانشجو:

فاطمه میرزاده

شماره دانشجویی: 40011920009

آذر 1400

خلاصه

با توسعه چشمگیر ارتباطات و محاسبات، کلاهبرداری بانکی در اشکال و مقادیر آن در حال رشد است. در این مقاله، ما اشکال مختلف کلاهبرداری را که بانک‌ها در معرض آنها قرار می گیرند و ابزارهای داده‌کاوی در معرض آن قرار می گیرند، تجزیه و تحلیل می‌کنیم که به داده‌های تشخیص زودهنگام آن که قبلاً در یک بانک انباشته شده است،منجر می شود. ما از روش‌های یادگیری نظارت‌شده از ماشین‌های بردار پشتیبانی با Spark (SVM-S) برای ساختن مدل‌هایی استفاده می‌کنیم که رفتار عادی و غیرعادی مشتری را نشان می‌دهند و سپس از آن برای ارزیابی اعتبار تراکنش‌های جدید استفاده می‌کنیم. نتایج به‌دست‌آمده از پایگاه‌های اطلاعاتی تراکنش‌های کارت اعتباری نشان می‌دهد که این تکنیک‌ها در مبارزه با کلاهبرداری بانکی در داده‌های بزرگ مؤثر هستند. نتایج آزمایش از این مطالعه نشان می‌دهد که SVM-S عملکرد پیش‌بینی بهتری نسبت به شبکه‌های انتشار برگشتی (BPN) دارد. علاوه بر میانگین پیش‌بینی، دقت زمانی که نسبت داده‌ها به 0.8 می‌رسد به حداکثر می‌رسد.

معرفی

سازمان های خدمات مالی دارای تعدادی اهداف استراتژیک از جمله جذب و حفظ مشتریان جدید و فعلی از طریق بکارگیری روش های مدیریتی مختلف هستند. با توجه به این اهداف، موسسات روزانه حجم زیادی از داده های پروفایل، تاریخچه خرید و مرور و داده های رسانه های اجتماعی را تولید می کنند. بانک‌ها به‌عنوان یک مؤسسه مالی، حجم عظیمی از داده‌ها را از منابع مختلف از مشتریان خود تولید می‌کنند و این به نیاز به داده‌های بزرگ طبق ویکی‌پدیا، (2016) کمک کرده است. ماشین بردار پشتیبانی می تواند به کاهش ریسک و بهبود کیفیت خدمات ارائه شده به مشتریان برای موفقیت در تجارت کمک کند. تقلب به یک خطر بسیار مهم برای مواجهه با موسسات مالی، اتحادیه های اعتباری و به ویژه بانک ها تبدیل شده است. مبارزه با کلاهبرداری با تکنیک‌های پیشگیری سنتی مانند پین، رمز عبور و سیستم‌های شناسایی همراه است، اما در سیستم‌های بانکی

مدرن ناکافی شده‌اند. بانک ها در چندین فعالیت با کلاهبرداری مواجه شدند اما به نظر می رسد استفاده از راه دور از اعتبار آسیب پذیرترین آنها باشد.

برنامه های کاربردی داده های بزرگ با تکنیک های داده کاوی می توانند نقش مهمی در مبارزه با این نوع تقلب ها داشته باشند. داده کاوی مجموعه ای از تکنیک ها برای استخراج اطلاعات مهم از مقادیر زیادی داده برای کمک به تصمیم گیری است Spark .یک ابزار کلان داده است که به‌ویژه در این زمینه استفاده می‌شود، زیرا به دلیل تکنیک‌های یادگیری ماشینی متعدد و جریان‌سازی در زمان واقعی است. شاغل شدن در موسسات مالی موثر است. جرقه با روش SVM برای این نوع کلاهبرداری ها بهترین است. چندین تکنیک برای کشف تقلب توسط بسیاری از محققین پیشنهاد شده و مورد استفاده قرار گرفته است، از جمله تقلب در کارت های اعتباری. از جمله این تکنیک‌های داده کاوی، شبکه‌های بیزی، زنجیره‌های مارکوف، شبکه‌های عصبی، رگرسیون خطی، هم‌ترازی توالی و غیره هستند. هدف این کار ارائه معماری تشخیص تقلب است که بانک را قادر می‌سازد تا تراکنش‌های متقلبانه را در زمان واقعی با جرقه بر اساس ماشین شناسایی کند. ماشین بردار پشتیبانی تکنیک یادگیری SVM .که برای تشخیص چهره، شناسایی اثر انگشت، تشخیص صدا و کارهای مشابه بسیار قدرتمند است. هدف ما رفع مشکل کلاهبرداری در بانک ها و حل آن از طریق Sparkبا تکنیک های SVM است. ما تحلیلی از مشکل کلاهبرداری بانک‌ها و برای هر نوع طراحی، گونه‌ای از SVM را ارائه می‌کنیم که می‌تواند برای راه‌حل آن و سازگاری‌های لازم استفاده شود.

بقیه مقاله به شرح زیر سازماندهی شده است: ابتدا انواع کشف تقلب و همچنین شاخص های مورد استفاده برای کشف آن ارائه می شود، سپس انواع راه حل های کلان داده را که می توان مورد استفاده قرار داد مورد بحث قرار می دهیم. در بخش سوم استفاده از ماشین بردار پشتیبان در جرقه برای رفع نیازهای تشخیص تقلب را مورد بحث قرار می دهیم. علاوه بر این، در بخش چهارم اعتبار سنجی راه حل های پیشنهادی را با آزمایش آنها بر روی پایگاه های داده بانک ارائه می کند. ما مقاله را با یک نتیجه گیری و توصیه به پایان می بریم.

II. اشکال کلاهبرداری بانکی و شاخص های آنها.

کلاهبرداری در بانک به اشکال مختلف ظاهر می شود. می تواند داخلی باشد که توسط کارمندان خود بانک متعهد شده باشد یا خارجی توسط مشتریان، افراد یا ارگان های خارجی به بانک متعهد شود. ما علاقه مندیم که در این مقاله تقلب خارجی را در نظر بگیریم.

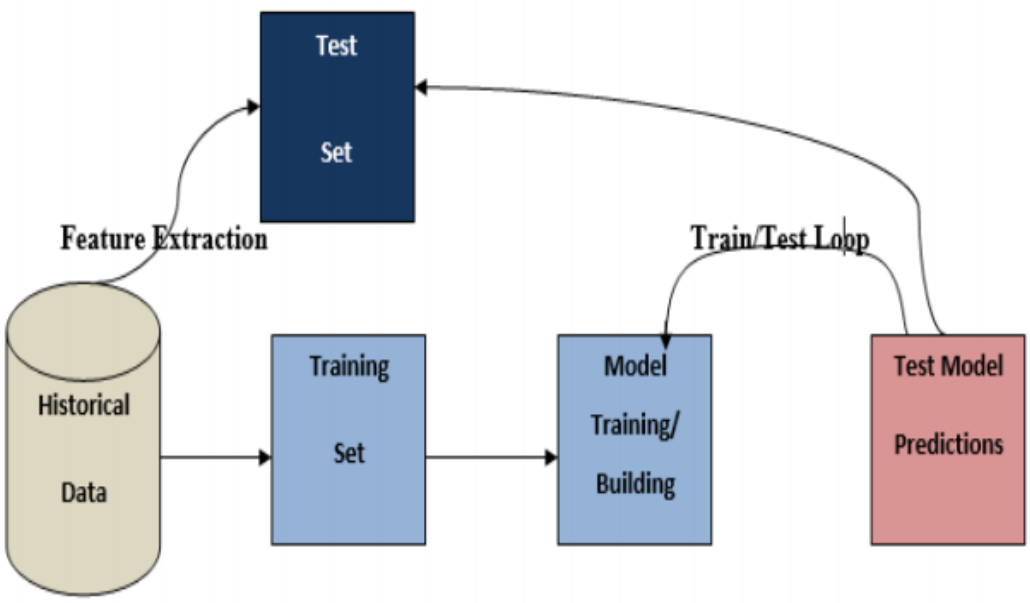
الف- پولشویی

پول شویی نیز یکی از روش های شناخته شده کلاهبرداری است، لوای بین المللی علیه این فعالیت توسط کشورهای مختلف برای کشف و پیگرد قانونی فعالیت های مجرمانه انجام می شود. مبارزه با پولشویی در صنعت مالی مبتنی بر تجزیه و تحلیل و پردازش اظهارات مربوط به تراکنش های مشکوک کشف شده توسط موسسات مالی است. به طور کلی، تنها چند تراکنش مشکوک واقعاً عملیات پولشویی هستند، اما تعداد عملیاتی که باید توسط مؤسسات مالی تجزیه و تحلیل شود، به زمان طولانی نیاز دارد. در ادبیات، روش‌های هوش مصنوعی ممکن است برای بهبود توانایی موسسات مالی در پردازش خودکار داده‌های مشکوک استفاده شود. با این حال، جستجو برای روش‌های کارآمد برای شناسایی رفتارهای مبادلاتی مشکوک پولشویی، یک زمینه تحقیقاتی بسیار فعال است. تعیین متغیرها و شاخص های پولشویی در حال حاضر آسان است، زیرا چنین فعالیت های غیر رسمی در شرایط پیچیده اقتصادی اجتماعی ایجاد می شوند. موارد زیر چند شاخص سفیدکننده پول مورد استفاده در ادبیات هستند. مبلغ تراکنش در صورتی که از مبلغ از پیش تعیین شده توسط بانک بیشتر باشد، معامله موجه نباشد، مشکوک است. به عنوان مثال، در غنا حداکثر مبلغ خرید با اعتبار شما نباید از 5000 یورو تجاوز کند. موارد زیر بیشتر مورد بررسی قرار می گیرد.

1. منابع انتقال
2. تاریخ معامله
3. تغییر آدرس
4. زمان معامله، معامله شبانه با مبلغ زیاد مشکوک است

ب. تقلب بر اساس کارت های اعتباری

کلاهبرداری مبتنی بر کارت اعتباری همچنان در حال افزایش است. بانک ها و شرکت های مالی سالانه مبالغ هنگفتی را از طریق کلاهبرداری با استفاده از کارت اعتباری از دست می دهند. تشخیص تقلب در کارت اعتباری اغلب بر اساس تعدادی از شاخص های پیش بینی است که عموماً از اطلاعات تراکنش های بازیابی شده از پایگاه داده تاریخی به دست می آیند. ما شاخص هایی مانند استفاده مکرر از کارت، مانده مانده پرداخت نشده هر چرخه، حداکثر تعداد روزهای تاخیر، دفعات خرید، تراکنش روزانه، بیشترین تعداد دفعات در پایگاه داده تاریخی و غیره را بررسی می کنیم. این ویژگی ها برای هر تراکنش استخراج می شوند و برای کشف الگوهای متقلبانه ثبت می شوندمدل تشخیص تقلب در شکل 1 نشان داده شده است.



شکل 1: مدل تشخیص تقلب

مدل داده پیشنهادی بر اساس داده‌های تاریخی موجود در انبار بانک، با استفاده از الگوریتم پیشنهادی که از SVM-S تشکیل شده است برای بررسی اینکه آیا نتیجه تقلبی است یا خیر، ساخته شده است. مجموعه ای از داده های مشابه برای پیش بینی نتیجه برای اثربخشی مدل استفاده خواهد شد. این مدل برای ارزیابی یک تراکنش جدید استفاده می شود. تراکنش پذیرفته شده توسط مدل اجرا می شود و سپس برای بهبود مدل به پایگاه داده اضافه می شود. تراکنش های رد شده

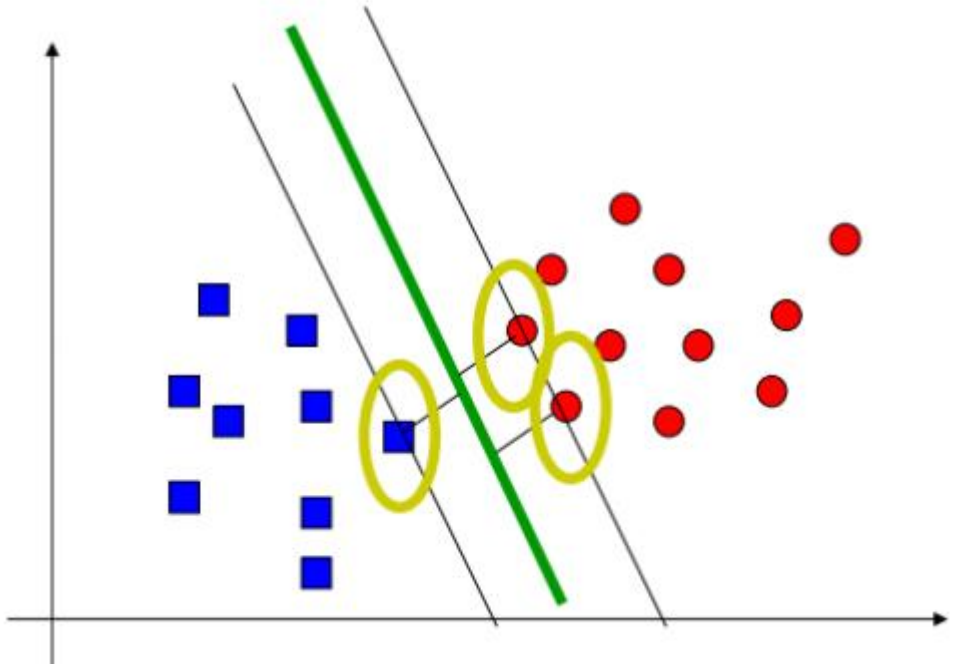
توسط مدل اجرا نمی شوند بلکه به عنوان مشکوک علامت گذاری می شوند. اگر تراکنش عادی باشد، همانطور که در بالا گفته شد اجرا و اضافه می شوند.

III. داده های بزرگ برای کشف کلاهبرداری بانکی

به دلیل تغییر سریع و توسعه تکنیک های مورد استفاده توسط کلاهبرداران، ابزارهای داده کاوی دیگر نمی توانند رفتارهای غیرعادی را تجزیه و تحلیل کنند. داده‌های بزرگ، در این زمینه، با تکنیک‌های یادگیری ماشینی برای تشخیص تقلب در پایگاه‌داده ارائه می‌شوند که بهترین کار برای مبارزه با کلاهبرداری بانکی است. در ادبیات، از دو شکل یادگیری خودکار استفاده می شود: نظارت شده و بدون نظارت. روشهای یادگیری نظارت شده به عنوان قوانین ارتباطی شبکه های بیزی استفاده شده است. این روش ها دانش قبلی از ماهیت معاملات، تقلبی یا واقعی را فرض می کنند. یادگیری در این مورد شامل ساخت مدلی است که فضا را با توجه به نمونه‌های موجود به دو قسمت تقسیم می‌کند و سپس نمونه‌های جدید را بر اساس عضویت آنها در یکی از این دو کلاس طبقه‌بندی می‌کند. روش های بدون نظارت مانند شبکه های عصبی نیازی به طبقه بندی قبلی از نمونه های آموزشی ندارد. بیشتر بر اساس تشخیص تراکنش های عجیب و غریب است. برنامه های کاربردی داده های بزرگ، مانندSpark ، Hadoop، Cassandraو غیره با الگوریتم های موثر برای مدیریت ساختار، داده های بدون ساختار و نیمه ساختار یافته ارائه می شوند.

الف. ماشین بردار پشتیبانی(SVM)

ماشین بردار پشتیبانی (SVM) یک الگوریتم یادگیری ماشینی نظارت شده است که می تواند برای چالش های طبقه بندی و رگرسیون استفاده شود. با این حال، بیشتر در مسائل طبقه بندی استفاده می شود. در این الگوریتم، ما هر آیتم داده را به عنوان یک نقطه در یک فضای بعدی رسم می کنیم)که در آن n تعداد ویژگی هایی است که شما دارید (با مقدار هر ویژگی مقدار یک مختصات خاص. شکل 2، فوق صفحه را نشان می دهد که این دو کلاس را متمایز می کند.



شکل 2: هایپر صفحه

ب. مورد خطی قابل تفکیک

در حالت قابل جداسازی خطی، یک یا چند ابرصفحه وجود دارد که ممکن است دو کلاس ارائه شده توسط داده های آموزشی را با 100٪ از هم جدا کنند

ج. مورد غیرخطی قابل تفکیک

در حالت غیرخطی قابل تفکیک، نمی توان یک ابر صفحه خطی پیدا کرد که تمام مثال های مثبت و منفی را از هم جدا کند. برای حل این مورد، تکنیک به حداکثر رساندن حاشیه ممکن است با اجازه دادن به برخی از نقاط داده در سمت اشتباه حاشیه، یعنی اجازه دادن درجه ای از خطا در جداسازی، تسهیل شود. متغیرهای Slack ε برای نشان دادن درجه خطا برای هر نقطه داده ورودی معرفی می شوند. ماشین بردار پشتیبانی برای طبقه بندی ویژگی های کارت های اعتباری توسط

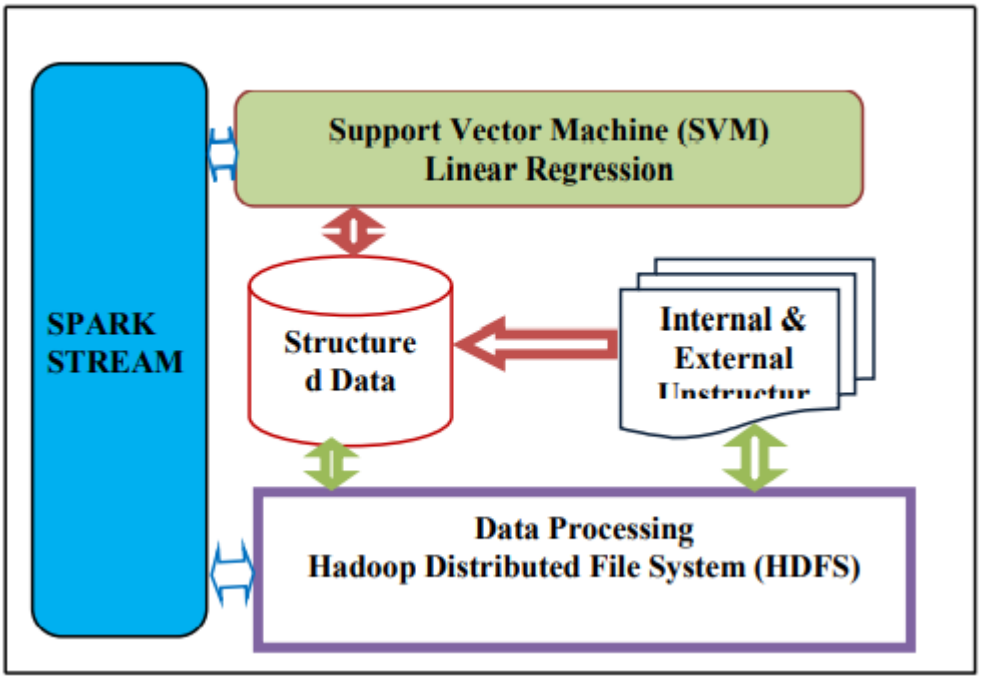
هر مشتری استفاده می شود SVM .باینری مشکل جداسازی دو کلاس را که با n مثال از هر کدام از ویژگی های m نشان داده شده است، حل می کند. مشکل زیر را در نظر بگیرید:

$$\{(x_1, y_1) \ldots\ldots\ldots (x_n, y_n)\} x_i \in \mathcal{R}m, y_i \in \{-1, +1\}$$

جایی که xi در حال یادگیری مثال ها و yi کلاس های مربوطه خود هستند. هدف روش SVM یافتن تابع خطی) f معادله 1) به نام hyperplane است که امکان جداسازی دو کلاس را فراهم می کند:

$$F(K) = (x \cdot w) + b$$

جایی که x نمونه ای برای طبقه بندی است، wیک بردار و b یک بایاس است. بنابراین ما باید وسیع ترین حاشیه را بین دو کلاس پیدا کنیم که معادل به حداقل رساندن است



شکل 3: معماری SVM-S برای تشخیص تقلب در داده های بزرگ

شکل 3، معماری SVM-S را برای چارچوب تشخیص تقلب نشان می دهد. برنامه spark همراه با پایگاه داده نیست. آن را همیشه در پایگاه داده خارجی برای عملیات آن سزاوار است HDFS .که در مدیریت کلان داده محبوب است، به عنوان پایگاه داده استفاده می شود. با تکنیک های یادگیری ماشین، ما SVM را برای آموزش مجموعه داده برای پیش بینی ها مستقر می کنیم.

2. Back Propagation Network:استفاده از BPN برای آموزش داده ها نیاز به تنظیم برخی پارامترها دارد. از جمله مهمترین پارامترها می توان به اعداد لایه پنهانNi ، گره پنهان Nj و دوره های آموزشیNk ، نرخ یادگیری Rj و نرخ تکانه Rm اشاره کرد. علاوه بر این، تنظیم مقادیر پارامتر به جای علم به عنوان یک هنر باقی می ماند. مشکلات پیچیده را می‌توان با افزایش لایه‌های پنهان به‌طور تدریجی بهتر مدل‌سازی کرد، اما این بهبود عموماً با هزینه‌های مرتبط از نظر زمان آموزش و تطبیق داده‌ها همراه است.

3.تست و نتیجه

آ. داده های مورد استفاده: به دلیل ماهیت محرمانه داده ها، دستیابی به داده های واقعی که رفتار مشتری بانک را توصیف می کند بسیار دشوار بود. با این حال، پایگاه‌های داده استانداردی وجود دارد که در ادبیات برای آزمایش روش‌های تشخیص تقلب استفاده می‌شود. برای آزمایش پیشنهاد هیبریداسیون، از پایگاه‌های داده آلمانی و استرالیایی کارت‌های اعتباری استفاده کردیم.

مجموعه داده برای پیاده سازی به حافظه برای پردازش فراخوانی شد. مجموعه داده پسوند فایل .csvبود. این داده‌های ایجاد شده به‌عنوان داده‌های تاریخی برای آموزش استفاده شد، یعنی با استفاده از الگوریتم نظارت‌شده. نام فایل برای مجموعه داده اول "creditcard.csv" و برای مجموعه داده دوم "creditcard.csv1" نامیده می شد. انتزاع اولیه Spark مجموعه ای توزیع شده از آیتم ها به نام مجموعه داده های توزیع شده انعطاف پذیر (RDD) است. در آن صورت، فایل

ها برای عمل و تبدیل به RDD تبدیل شدند. مجموعه داده با استفاده از SVM طبقه‌بندی شد و با رگرسیون خطی و رگرسیون منطقی برای تشخیص ناهنجاری با استفاده از ویژگی‌های کارت اعتباری آموزش دید. ترکیب این سه روش در مدل ما باعث افزایش دقت تشخیص می شود. جدول 1 پارامترهای تنظیم شده برای مجموعه داده را نشان می دهد.

Data Set	Classifier Model	Parameters	Definition	Source
1st Data Set	SVM	C=10 RBF ε = 0.1		Djeffal et al. 2014
2nd Data Set	BPN	C=10 RBF ε = 0.1		Soltani et al. 2014

جدول 1: پارامترهای ورودی برایSVM

مدل SVM-S ساخته شده برای دو مجموعه داده از یک تابع هسته استفاده می کند. دو پارامتر مرتبط با هسته RBF وجود دارد C :و .ε پارامتر تنظیم (C) مبادله بین حداکثر کردن حاشیه و به حداقل رساندن مدت خطای آموزشی را کنترل می کند. افزایش مقدار باعث بهبود دقت طبقه بندی (کاهش خطای رگرسیون) برای آموزش داده می شود. برای یکSVM-S ، مقدار ε در تابع ضرر غیر حساس به ε نیز باید انتخاب شود ε .روی صاف بودن پاسخ SVM تأثیر می گذارد و بر تعداد بردارهای پشتیبانی تأثیر می گذارد، بنابراین هم پیچیدگی و هم قابلیت تعمیم شبکه به مقدار آن بستگی دارد.

Samples		# of Records	
		Training Set	Test Set
Set -1F-To-1N (SVM-S)	Normal	100	86
	Fraud	25	20
Set -1F-To-1N (BPN)	Normal	150	125
	Fraud	25	12

جدول 2: اندازه مجموعه آموزش و آزمون برای نمونه ها

دو مجموعه داده مختلف برای آزمایش هر مدل الگوریتم، با شرایط مختلف استفاده شد. اولین مجموعه داده یک تراکنش عادی برای هر تراکنش متقلبانه دارد. در حالی که مجموعه دوم، دارای چهار نرمال برای هر یک تقلبی است. برای هر داده نمونه، 70 درصد از داده ها، 70 درصد تراکنش عادی و 70 درصد جعلی، به عنوان مجموعه آموزشی برای مدل در نظر گرفته می شود. در حالی که 30 درصد از داده ها به عنوان مجموعه آموزشی برای ارزیابی عملکرد مدل مستقر شده در نظر گرفته می شود. اندازه مجموعه آموزش و آزمون در جدول 2 در بالا آورده شده است.

ب نتیجه: عملکرد معماری SVM-S در جدول 3 ارائه شده است، ستون سمت چپ روشی که برای ساخت مدل های معماری استفاده شده است، ستونی با نام "train" نشان دهنده دقت پیش بینی معماری پیشنهادی بر روی مجموعه داده های آموزشی است. در نمونه های داده شده، ستون هایی با برچسب "تست" دقت پیش بینی مدل معماری را در مجموعه داده های آزمایشی نمونه های داده شده نشان می دهد. ستونی که به عنوان "Build Time" نامگذاری شده است، زمان سپری شده برای ساخت مدل معماری داده شده را در نمونه داده شده نشان می دهد و ستون های "Frauds" تعداد تراکنش های تقلبی را در مجموعه داده های اختصاص داده شده به عنوان تقلب (TruePositive)توسط معماری بر روی نمونه های داده شده نشان می دهد. .

مقایسه نتایج پیش‌بینی‌شده بین BPN و SVM-S جدول 3، مقایسه نتایج پیش‌بینی‌شده بین BPN و SVM-S را از آزمایش نشان می‌دهد.

Model/ Data Set	Set-1F-To-1N				Set-1F-To-4N			
	Train	Test	Build Time	Fraud	Train	Test	Build Time	Fraud
D1: SVM	98.78%	84.37%	<20m	20	96.34%	82.54%	<20	20
D2: BPN	99.86%	85.58%	<25m	13	97.346%	84.67%	<25	10

جدول 3: دقت عملکرد نسبت به مجموعه های آموزشی و آزمایشی

از جدول 3، مشخص است که معماری SVM-S با عملکرد SVM بسیار کارآمد است. با این حال، دقت میزان تخصیص واقعی را بدون توجه به اینکه یک تقلب واقعی است یا انتساب عادی واقعی، نشان می دهد. با این وجود، تعداد تراکنش‌های متقلبانه که توسط مدل به عنوان جعلی در نظر گرفته شده بود، برای مدل دقیق بود. این واقعیت که نتیجه مورد انتظار تعداد تراکنش های تقلبی تزریق شده با هر مجموعه داده، نتیجه SVM-S برابر است با کار در مقایسه با نتیجه .BPN این به وضوح نشان می دهد که SVMS از BPN قابل اعتمادتر و دقیق تر است. با پیچیدگی زمانی، SVM-S از زمان کمی برای پیش‌بینی ناهنجاری‌ها در مقایسه با الگوریتم‌های دیگر مانند BPN استفاده می‌کند.

IV. نتیجه

در این زمینه دو مورد کلاهبرداری در بانک‌ها را بررسی کردیم: کلاهبرداری از کارت اعتباری و پولشویی. عملکرد سیستم پیشنهادی بر روی معیارهای دفتر کل، داده های پرداختنی، که شبیه به

پایگاه داده بانک ایجاد شده است، آزمایش شد. دقت به‌دست‌آمده برای روش SVM تک کلاس حدود 80 درصد بود که نشان‌دهنده پیشرفت قابل‌توجهی در مقایسه با مرجع آثار مشابه است. برای این روش، بهبود جزئی در پایگاه‌های داده امتیازدهی اعتباری به دلیل دشواری دستیابی به پایگاه‌های داده واقعی بود. نتایج را می توان با مطالعه تأثیر پارامترهای مختلف مورد استفاده در معماری SVM-Sبهبود بخشید.