ارائه مدل پیاده سازی سیستم کشف تقلب کارت‌های اعتباری در محیط ابری با استفاده از نگاشت-کاهش

ندا سلطانی حلوایی[۱]، محمدکاظم اکبری[۲]، مرتضی سرگلزایی جوان[۳]
[۱]دانشگاه صنعتی امیرکبیر، neda.soltani@aut.ac.ir
[۲]دانشگاه صنعتی امیرکبیر، akbarif@aut.ac.ir
[۳]دانشگاه صنعتی امیرکبیر، msjavan@aut.ac.ir

چکیده - کشف تقلب یکی از فعالیت‌های حیاتی در بانک‌ها و سازمان‌های صدور کارت‌های اعتباری است. علیرغم تمهیدات امنیتی تقلب به سرعت در حال گسترش است. استفاده از روش‌های پرداخت آنلاین و تلفنی یکی از دلایل این امر است. بنابراین لزوم استفاده از روش‌های کشف و جلوگیری از ثبت تقلب علاوه بر اقدامات پیش‌گیرانه، بدیهی است. عملیات کشف یک تراکنش مشکوک و تصمیم گیری در مورد آن باید حتی الامکان قبل از ثبت تراکنش انجام شود لذا همه پردازش در زمان کوتاهی باید صورت گیرد. تعداد تراکنش‌های ثبت شده در سیستم در طول یک روز زیاد است و لازم است این تراکنش‌ها به شکل آنلاین مدیریت و بررسی شوند. در این مقاله کشف تقلب کارت‌های اعتباری هدف قرار داده شده است. به دلیل شباهت موجود بین سیستم کشف تقلب و سیستم ایمنی مصنوعی از این سیستم به عنوان روش کشف استفاده شده است. الگوریتم انتخاب شده یک الگوریتم دسته بندی نظارت شده است که زمان زیادی برای آموزش نیاز دارد. برای کاهش زمان آموزش الگوریتم و دستیابی به سریع‌ترین پاسخ ممکن، از رایانش ابری به عنوان محیط پیاده‌سازی استفاده شده است. بدین ترتیب زمان آموزش الگوریتم به شکل قابل توجهی کاهش یافته است و می‌توان کشف کننده‌های جدید را به سرعت به سیستم اضافه کرد و بروزرسانی را در حداقل زمان انجام داد. نتایج نشان می‌دهد که زمان آموزش الگوریتم به شکل قابل توجهی کاهش یافته است.

کلید واژه‌ها- کشف تقلب، رایانش ابری، نگاشت/کاهش، سیستم ایمنی مصنوعی.

۱. مقدمه

تقلب صورت گرفته در حوزه کارت‌های اعتباری، باعث تحمیل هزینه به بانک‌ها و سازمان‌های صدور کارت‌های اعتباری می‌شود و شهرت آن‌ها را به خطر می‌اندازد. متمرکز شدن روی انواع روش‌های کشف تقلب و بررسی روش‌های جدید برای مقابله و پیشگیری از این موارد، اهمیت بالایی دارد. از این رو امروزه کشف تقلب یک فعالیت حیاتی تجاری برای کمینه کردن تاثیرات تراکنش‌های غیرمجاز روی تحویل سرویس به مشتری، هزینه‌ها و شهرت تجاری یک سازمان است؛ که از طریق به کار گرفتن چارچوب‌های مبتکرانه تکنولوژی کشف تقلب صورت می‌گیرد. هرچه سازمان‌ها برای دفاع در برابر روش‌های شناسایی شده، تمهیدات پیشرفته تری اتخاذ کنند، روش‌های جدیدتری برای تقلب گسترش می‌یابد. برای کشف و ممانعت از تقلب تکنولوژی‌های متعددی مورد استفاده قرار می‌گیرد. مکانیزم‌های کشف تقلب تراکنش‌های تقلبی انجام شده از طریق رخنه‌های امنیتی را تشخیص می‌دهند؛ در حالی که سطوح بهینه ارائه سرویس حفظ می‌شود و تعداد هشدارهای اشتباه کمینه نگه داشته می‌شود. مرجع [۱] اضافه شدن هزینه‌های وابسته به تقلب را دلیل گرایش به سمت روش‌های پیش گیرانه کشف تقلب معرفی کرده است؛ بدین معنی که بتوان آزمایش داده‌های تراکنشی را به شکل بلادرنگ و کشف رفتارهای مبهم کاربر را قبل از تکمیل تراکنش انجام داد. این انتقال پردازش داده از "بعد" به "قبل" از ذخیره‌سازی داده، به شکل قابل توجهی زمان مقدور را برای ارزیابی تقاضاهای جدید از سیستم و اتخاذ یک تصمیم دقیق برای تقلب کاهش می‌دهد. روزانه ده‌ها هزار تراکنش توسط دستگاه‌های POS، درگاه‌های پرداخت آنلاین، خودپردازها و ... به سمت بانک‌ها سرازیر می‌شود که هر یک از این تراکنش‌ها می‌تواند مجاز یا تقلبی باشد. علاوه بر اینکه پردازش هر یک از این تراکنش‌ها عملی زمان‌گیر است، لازم است رفتارهای قبلی کاربران در نظر گرفته شده و سیستم بر اساس

تراکنش‌های قبلی آموزش یابد؛ که این امر هم به نوبه خود زمان زیادی لازم دارد. بر همین اساس نیاز به بستری قدرتمند داریم تا با حداقل کردن زمان پردازش تراکنش‌ها و نیز زمان آموزش الگوریتم و از طرفی بیشینه کردن دقت الگوریتم، به بهترین نحو و به شکل بلادرنگ از انجام تقلب جلوگیری کند. رایانش ابری چنین بستری را فراهم می‌کند. در بستر ابر می‌توان علاوه بر دسترسی به قدرت بالای پردازشی و محاسباتی، به مجموعه وسیعی از تراکنش‌ها و اطلاعات به اشتراک گذاشته شده در مورد تقلب، توسط سازمان‌های مختلف دسترسی داشت که تا حد زیادی به پوشش دادن انواع مختلف تقلب و بالا بردن دقت الگوریتم کشف تقلب کمک می‌کند. ادامه این مقاله بدین شکل است: در قسمت ۲ معرفی مختصری در مورد سیستم ایمنی مصنوعی و الگوریتم استفاده شده، ارائه شده است. در بخش ۳ مدل پیشنهادی برای سیستم ارائه شده است و در قسمت ۴ معماری مدل پیشنهادی توضیح داده شده است. در نهایت قسمت ۵ شامل نتایج است.

۲. سیستم ایمنی مصنوعی

در حوزه روش‌های کشف تقلب، تا کنون تحقیقات زیادی صورت گرفته است. یکی از روش‌های کشف تقلب الهام گرفته از سیستم ایمنی بدن انسان است و با مبنا قرار دادن مشابهت‌های سیستم ایمنی بدن با سیستم کشف تقلب، می‌تواند بهبود مناسبی در حوزه کشف تقلب ارائه دهد. سیستم ایمنی مصنوعی که از سیستم ایمنی بدن انسان الهام گرفته شده است، تشخیص سلول‌های غیرخودی از سلول‌های خودی را هدف قرار می‌دهد. روش کار آن شامل تولید کشف‌کننده‌هایی است که در برابر سلول‌های خودی مقاوم‌اند و سلول‌های غیرخودی را شناسایی می‌کنند. این سیستم قابلیت یادگیری انواع جدید سلول‌های غیرخودی در طول عمر سیستم را داراست. همچنین تطبیق‌پذیر است و می‌تواند در طول زمان انواع جدید سلول‌های غیرخودی را تشخیص دهد و در حافظه‌اش ذخیره کند.

۲,۱. انتخاب کلونی

یکی از پروسه‌های اصلی در سیستم ایمنی مصنوعی، انتخاب کلونی است. سلول های کشف کننده به نحوی تولید می‌شوند که با سلول‌های خودی بدن واکنش نمی‌دهند، بنابراین زمانی که با سلولی واکنش دهند –به اصطلاح شباهت بالایی با سلول مزبور داشته باشند- احتمال اینکه سلول مورد آزمایش غیرخودی باشد بالا است. پروسه انتخاب کلونی برای اطمینان از غیرخودی بودن سلول مورد آزمایش انجام می‌گیرد. پروسه انتخاب کلونی زمانی اتفاق می‌افتد که یک سلول کشف کننده، سلول غیرخودی‌ای را تشخیص دهد. در این حالت کشف کننده با تقسیم سلولی تکثیر می‌شود. این پروسه شامل دو بخش است که از نظر محاسباتی دارای اهمیت است: اول اینکه هر سلول غیرخودی می‌تواند تعداد زیادی از سلول‌های کشف‌کننده را برای تکثیر تحریک کند. دوم اینکه میزان تکثیر با درجه شباهت رابطه مستقیم دارد. این تکثیر با جهش یا تغییر همراه است. هر چه میزان شباهت بیشتر باشد، میزان تغییر کمتر خواهد بود. این جهش کمک می‌کند کشف کننده‌های دقیق‌تری تولید شوند. الگوریتم انتخاب کلونی برای کشف تقلب در ادمه آمده است. لازم به ذکر است که در سیستم کشف تقلب مبتنی بر انتخاب کلونی، آموزش بر اساس رکوردهای تقلب برچسب خورده انجام می‌شود. بنابراین زمانی تکثیر انجام می‌شود که شباهت بین کشف‌کننده و رکورد تقلب از حد آستانه بالاتر باشد.

الگوریتم انتخاب کلونی [۲]:

۱. کشف‌کننده‌ها به صورت تصادفی تولید می‌شوند.
۲. برای هر جفت کشف‌کننده و رکورد داده تقلبی، درجه شباهت محاسبه می‌شود.
۳. کشف‌کننده‌هایی که میزان شباهت آن‌ها با رکورد مورد آموزش بالاست، انتخاب و تکثیر می‌شوند.
۴. بسته به میزان شباهت، این کشف‌کننده‌ها جهش می‌یابند.
۵. کشف‌کننده‌هایی که بیشترین شباهت را داشته‌اند، به عنوان سلول حافظه انتخاب می‌شوند.
۶. مراحل دوم تا پنجم آنقدر تکرار می‌شوند تا شرط خاصی برآورده شود؛ مثلا تعداد کشف کننده‌های مورد نظر.

پس از آموزش سیستم بر اساس رکوردهای تقلبی، سلول‌های حافظه تولیدشده برای دسته بندی رکوردهای تست مورد استفاده قرار می‌گیرند.

۲,۲. الگوریتم تشخیص ایمنی مصنوعی

الگوریتم انتخابی در این مقاله، الگوریتم تشخیص ایمنی مصنوعی[1] است که بر اساس انتخاب کلونی عمل می‌کند و در [۳] معرفی شده است. نحوه کار الگوریتم بدین شکل است که وقتی که شباهت بین یک سلول کشف‌کننده و یک سلول غیرخودی بالا باشد، سلول کشف‌کننده وارد پروسه انتخاب کلونی می‌شود که به نسبت میزان شباهتش با غیرخودی تکثیر می‌شود. برای محاسبه شباهت از فاصله اقلیدسی استفاده شده است. پس از آن پروسه بلوغ شباهت[2] صورت می‌گیرد که در آن جهش با نسبت عکس با شباهت انجام می‌شود. در نهایت در یک پروسه دیگر، سلول‌های حافظه انتخاب می‌شوند. در AIRS این سلول‌های حافظه برای دسته‌بندی استفاده می‌شوند. دسته بندی از طریق الگوریتم kNN صورت می‌پذیرد. هر عنصر داده‌ای در مجموعه تست، به هر سلول حافظه ارائه می‌شود و همسایگی آن‌ها که معادل میزان تحریک است، محاسبه می‌شود. الگوریتم kNN از این مقدار همسایگی برای دسته‌بندی استفاده می‌کند.

۲,۳. چالش‌های روش

AIRS نتایج خوبی در مقایسه با روش های کشف تقلب دیگر نظیر درخت رگرسیون، شبکه های عصبی و شبکه های بیزین دارد. با تنظیم مناسب پارامترهای این الگوریتم می تواند به دقت بالایی در کشف تقلب دست یافت [4]. اما الگوریتم مزبور پردازش زیادی در مرحله آموزش انجام می‌دهد. تولید یک سلول حافظه مستلزم محاسبه فاصله سلول مورد آموزش از همه سلول‌های حافظه و همه جهش‌های تولید شده از آن در دفعات مکرر است. ضمن اینکه برای بدست آوردن حد آستانه فاصله در اولین مرحله آموزش فاصله بین هر دو سلول در مجموعه آموزش محاسبه می‌شود. هرچه تعداد رکوردهای آموزشی، یا تعداد سلول‌های حافظه مورد نیاز بیشتر باشد، این زمان هم زیاد می‌شود. اهمیت این امر در محیط واقعی مشخص می‌شود چرا که همچنان که پیش‌تر نیز اشاره شد، تعداد رکوردهای ثبت شده زیاد است و دستیابی به دقت عمل بالا در سیستم، مستلزم در نظر گرفتن تراکنش‌های ثبت شده در دوره مشخصی است.

۳. متدولوژی

در این مقاله مدلی برای پیاده‌سازی سیستم کشف تقلب بر اساس نگاشت-کاهش[3] ارائه می‌شود که تا حد زیادی مشکل زمان مورد نیاز آموزش سیستم تشخیص ایمنی مصنوعی را کاهش می‌دهد. این مدل روی سیستم فایل Hadoop عمل می‌کند.

۳,۱. پیاده سازی در بستر ابر

رایانش ابری علاوه بر ارائه قدرت محاسباتی بالا که پردازش حجم بالای تراکنش‌ها را ممکن می‌کند -این بدلیل زیرساخت الاستیک ابر است- کاهش هزینه‌ها را نیز در پی دارد چرا که به دلیل برون سپاری هزینه‌های کشف تقلب کاهش می‌یابد. همچنین در محیط ابری دسترسی به منابع زیرساخت به میزان نیاز است که این امر سازمان را از نگرانی در مورد مدیریت زیرساخت رها می‌کند.

مدل کشف تقلب پیشنهادی از بستر ابر برای اجرا استفاده می‌کند. این مدل با در دست داشتن داده‌های سازمان‌های مختلف می‌تواند الگوریتم کشف تقلب را روی آن‌ها اعمال کند و مدل‌های تقلب را استخراج کند. زمان لازم برای آموزش الگوریتم‌های زیستی نسبتاً زیاد است و پردازش زیادی روی تک تک رکوردها صورت می‌گیرد. در الگوریتم AIRS دو مرحله زمان‌گیر وجود دارد: اول محاسبه حد آستانه شباهت که شامل محاسبه شباهت میان هر دو عنصر در مجموعه داده آموزش و میانگین‌گیری از آن است. دوم تولید سلول‌های حافظه که روی تک تک رکوردهای آموزش پردازش انجام می‌شود. به دلیل که لزوم به روزرسانی مکرر سیستم، کاهش زمان اجرای الگوریتم دارای اهمیت است؛ مخصوصاً اگر تعداد رکوردهای آموزشی - یعنی تعداد تراکنش‌های ثبت شده که باید پردازش شوند- زیاد باشد. یکی از روش‌های کاهش زمان، استفاده از موازی‌سازی است.

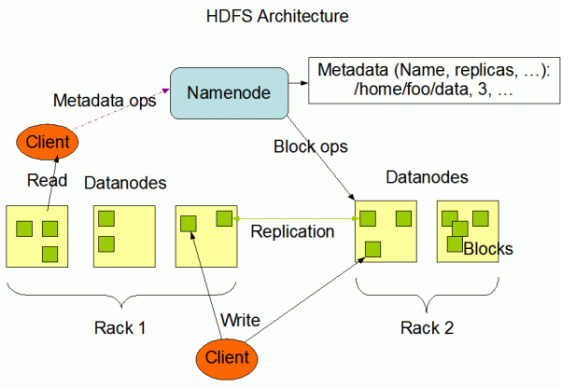
[1] Artificial Immune Recognition System (AIRS)
[2] Clonal
[3] Map-Reduce

۳,۲. مدل مبتنی بر نگاشت-کاهش

راهکاری که برای کاهش زمان در پروسه های زمان گیر روی حجم بالای داده ها پیشنهاد می شود، موازی سازی است. موازی‌سازی در مواجهه با داده‌های حجیم اهمیت بیشتری می‌یابد. در تئوری راه حل آسان است: می‌توان داده‌ها را در دسته‌های مساوی در پروسه‌های متفاوت با استفاده از همه امکانات سخت افزای موجود، پردازش نماییم. با این حال در عمل مشکلاتی وجود دارد:

- تقسیم کار به بخش‌های مساوی همیشه آسان و بدیهی نیست.
- ترکیب نتایج پروسه‌های وابسته نیازمند پردازش اضافی است.
- پردازش محدود به قدرت و زمان پردازش ضعیف‌ترین ماشین است وممکن است داده‌ها بزرگتر از ظرفیت ماشین باشند.
- لازم است ماشین‌ها و نیز پروسه‌های ناموفق به شکل متمرکز مدیریت شوند.

بنابراین اگرچه استفاده از موازی‌سازی ممکن است، در عمل دشواری‌هایی دارد. موازی‌سازی روی بستر ابر مشخصه‌های مطلوبی دارد که کمک بیشتری در این پروژه می‌کند. اول اینکه محدودیت‌های ماشین‌ها در زمان و فضای مورد نیاز دخیل نمی‌شود و دیگر اینکه مدیریت اجرای پروسه‌های موازی به صورت خودکار انجام می‌شود. در این مقاله پیاده سازی در بستر ابر با استفاده از نگاشت-کاهش که یک واسط برنامه نویسی کاربردی است، صورت گرفته است. این واسط از Apache Hadoop که یک سیستم فایل توزیع شده روی ابر است، استفاده می‌کند. Hadoop چارچوبی برای اجرای برنامه‌ها روی کلاستر است که قابلیت اطمینان و حرکت داده‌ها را فراهم می‌کند. Hadoop از پارادایم محاسباتی نگاشت-کاهش استفاده می‌کند که برنامه را به قسمت‌های کوچک‌تر تقسیم و روی نودهای مختلف کلاستر اجرا می‌کند. علاوه بر آن Hadoop از یک فایل سیستم توزیع شده (HDFS)[4] استفاده می‌کند که داده را روی نودهای محاسباتی ذخیره می‌کند (شکل ۱). هر دوی این بخش‌ها طوری طراحی شده اند که نودهای ناموفق به شکل اتوماتیک توسط چارچوب مزبور مدیریت می‌شوند.



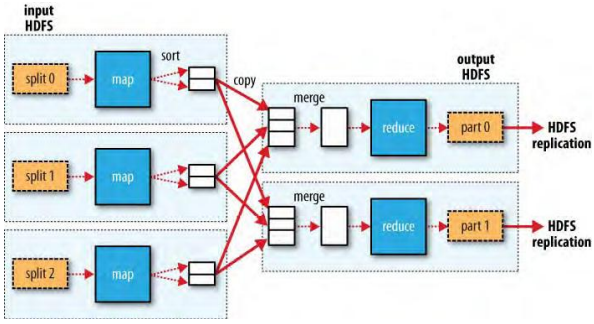
شکل ۱: معماری HDFS [۵]

استفاده از چارچوبی نظیر Hadoop چالش‌های مزبور را حل می‌کند:

- کارها به شکل اتوماتیک تقسیم می‌شوند.
- نتایج بدست آمده توسط مکانیزم "کاهش" ترکیب می‌شوند. ضمن اینکه عملیات مورد نیاز برای ترکیب نتایج، توسط برنامه نویس قابل تعریف است.
- بدلیل مجازی بودن ماشین‌ها محدودیت قدرت ماشین‌ها مطرح نیست.
- بدلیل توزیع شده بودن سیستم فایل، محدودیت فضای ذخیره سازی ماشین‌ها مطرح نیست.
- مدیریت متمرکز توسط ماشین master انجام می‌شود در حالیکه کارهای واگذار شده به ماشین‌ها به شکل مستقل اجرا می‌شود.

نگاشت-کاهش با استفاده از تقسیم کار به دو فاز نگاشت و کاهش کار می‌کند. برای استفاده از ساختار نگاشت-کاهش باید دو تابع نگاشت و کاهش تعریف شوند [۶]. در واقع اجرای یک برنامه با استفاده از این چارچوب نیازمند پیاده سازی بخش‌های موازی تحت تابع نگاشت است. می‌توان از تابع کاهش برای جمع بندی نتایج کارهای موازی و بدست آوردن خروجی نهایی استفاده کرد. شکل ۲ عملکرد این چارچوب را نشان می‌دهد. تابع نگاشت روی هر بلوک ورودی فراخوانی می‌شود.

[4] Hadoop Distributed File System



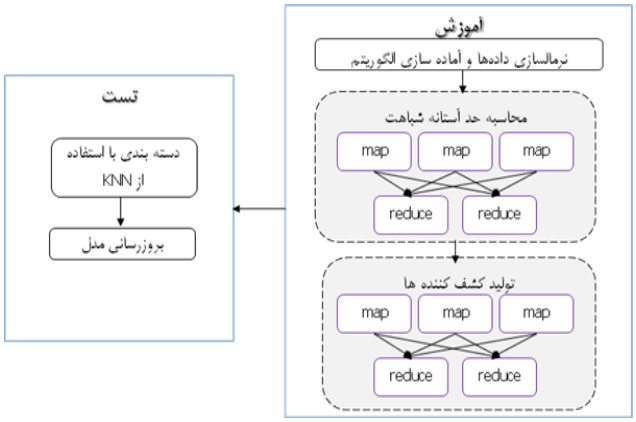
شکل ۲: جریان داده در نگاشت-کاهش [۶]

۳,۳. به روز رسانی مدل

یکی از ویژگی‌های مطلوب یک سیستم کشف تقلب بروز بودن آن است. سیستم باید بتواند انواع جدید تقلب را یاد بگیرد و این یادگیری باید در طول کار سیستم ادامه داشته باشد. بر همین اساس در محیط‌های آزمایشگاهی، سیستم کشف تقلب می‌تواند در حین بررسی مجموعه داده تست، به‌روزرسانی شود و مرحله آموزش را روی رکوردهای تست تکرار کند. در مدل سیستم کشف تقلب پیشنهادی، در مرحله تست رکوردها، هر رکورد پس از دسته بندی وارد مرحله آموزش می‌شود تا سلول حافظه مناسب برای آن تولید شود. این سلول حافظه به مدل اضافه می‌شود و در بررسی رکوردهای بعدی مورد استفاده قرار می‌گیرد. برای شبیه‌تر شدن مدل به محیط واقعی، داده‌ها بر اساس زمان ثبت تراکنش مرتب شده‌اند. ضمن اینکه به روزرسانی در دوره‌های متناوب، پس از تست ۱۰۰ رکورد اجرا می‌شود تا نتایج واقعی‌تر باشد و این مفهوم که قطعی شدن تقلبی بودن یک تراکنش در محیط واقعی زمان‌بر است، پوشش داده شود.

۳,۴. مدل پیشنهادی

معماری مدل نهایی بر اساس آنچه که در بخش‌های پیشین ارائه شد، در شکل ۳ نمایش داده شده است. در بخش آموزش پس از نرمالسازی داده‌ها، بخش اول نگاشت برای محاسبه حد آستانه شباهت اجرا می‌شود و فاصله هر دو رکورد در مجموعه داده آموزش محاسبه می‌شود. در مرحله کاهش میانگین فاصله محاسبه شده و در خروجی ثبت می‌شود. در ادامه فاز اصلی آموزش و تولید کشف‌کننده‌ها با استفاده از تابع نگاشت آغاز می‌شود. تابع نگاشت داده‌های ورودی را بین توابع کاهش تقسیم می‌کند و کشف‌کننده‌ها در تابع تقسیم تولید شده و در نهایت در خروجی ثبت می‌شوند. تولید کشف‌کننده‌ها شامل انتخاب کلونی، انتخاب منفی و رتبه دهی کشف‌کننده‌های تقلب است. با توجه به اینکه هریک از توابع کاهش روی بخشی از داده‌های آموزشی کار می‌کنند، ممکن است تولید کشف‌کننده‌ها همراه با افزونگی‌هایی باشد به این معنی که برخی سلول‌های حافظه روی چندین گره تولید شوند. باید توجه داشت این امر به افزایش دقت کمک می‌کند. آزمایش‌های انجام شده نشان می‌دهد افزایش تعداد کشف‌کننده‌ها باعث افزایش دقت می‌شود. در مرحله کشف پس از دسته‌بندی رکوردها، به‌روزرسانی مدل انجام می‌شود.



شکل ۳- مدل ارائه شده برای سیستم کشف تقلب در محیط رایانش ابری

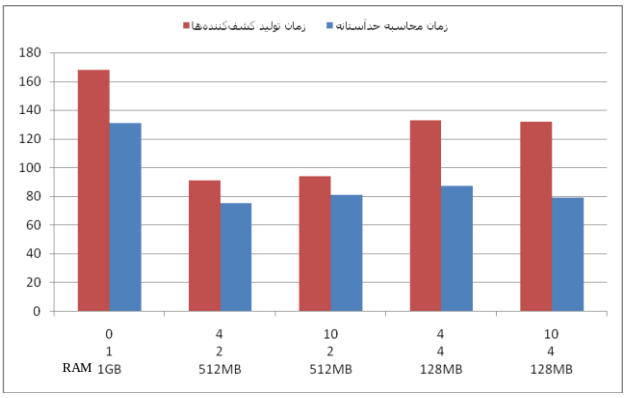
۴. نتایج

در این بخش نتایج مربوط به تست مدل معرفی شده ارائه می‌شود. برای ارزیابی بروزرسانی مدل، علاوه بر نرخ کشف، نرخ مثبت اشتباه و نسبت مثبت اشتباه به مثبت درست، از معیاری که در مقاله قبلی نویسنده [۷] بر اساس هزینه معرفی شده است نیز استفاده شده است. برای ارزیابی پیاده سازی روی محیط ابری، از مقایسه زمان اجرای الگوریتم استفاده شده است که نشان دهنده کاهش این زمان در صورت اجرا روی ابر است.

ارزیابی بر اساس مجموعه داده‌ای صورت گرفته است که مربوط به یک صادرکننده کارت اعتباری است و دارای ۴۱۶۴۷ رکورد با ۳,۷۴٪ تقلب است که در پنجره زمانی ۱۴ جولای ۲۰۰۴ تا ۱۲ سپتامبر ۲۰۰۴ ثبت شده‌اند. این مجموعه داده

برای تست در مرجع [۴] استفاده شده است. فیلدهای رکوردها برای حفظ حریم خصوصی دسته‌بندی شده‌اند. داده‌ها به فرمت Weka هستند[5] که الگوریتم AIRS به شکل یک افزونه به آن اضافه شده است. انواع مختلف تقلب در این مجموعه داده وجود دارد، شامل: کارت‌های گمشده یا دزدیده شده، دزدی اطلاعات کارت توسط فروشنده[6]، سفارشات تلفنی و ایمیل، کنترل حساب[7].

مدل ارائه شده روی محیط ابری پیاده‌سازی و تست شده است. نتایج تست و مشخصه‌های ماشین‌های مجازی که تست روی آن‌ها انجام شده است در جدول ۱ آمده است. ستون "تعداد توابع نگاشت" نشان دهنده تعداد توابع نگاشتی است که در بخش موازی‌سازی شده الگوریتم (محاسبه حد آستانه شباهت و تولید کشف‌کننده‌ها) وجود دارد. نتایج بر اساس تغییر تعداد توابع نگاشت و نیز تغییر تعداد گره‌ها ثبت شده است. با توجه به اینکه در هر دو بخش موازی سازی شده توابع کاهش فقط وظیفه جمع‌آوری نتایج توابع نگاشت را دارند، تغییر تعداد آن‌ها تاثیر زیادی در زمان اجرای الگوریتم ندارد. اما افزایش تعداد توابع نگاشت در صورتی که تعداد رکوردهای داده زیاد باشد باعث کاهش زمان اجرا می‌شود. زمان‌های هر بخش بر اساس ثانیه آمده است. اگرچه از ماشین‌های مجازی برای پیاده سازی Hadoop استفاده شده است، کاهش زمان مشهود است. استفاده از تعداد بیشتر ماشین‌های مجازی مستلزم کاهش منابع هر گره است. بنابراین در مرحله تولید کشف کننده‌ها، زمان قدری افزایش یافته است. با مقایسه تست‌های شماره ۲ و۳ می‌بینیم که افزایش تعداد توابع نگاشت تاثیر چندانی ندارد. چرا که تعداد گره‌ها ثابت است و موازی سازی فقط روی دو تابع نگاشت به شکل همزمان صورت می‌گیرد. با این حال زیاد شدن تعداد ماشین‌های مجازی که موازی سازی تعداد بیشتری تابع نگاشت را ممکن می‌کند، در تعداد بالای رکوردهای آموزشی نتایج بهتری خواهد داشت. همچنین نمودار ۱ تغییرات زمان را برای هر دو مرحله موازی سازی شده نمایش می دهد. جدول ۲ نشان دهنده نتایج حاصل از اضافه کردن بخش به روز رسانی به سیستم است. همچنانکه مشخص است دقت الگوریتم با اضافه شدن بخش به روز رسانی بیشتر شده است. این امر به دلیل اضافه شدن انواع جدید تقلب به کشف کننده ها است.



نمودار ۱- نتایج پیاده سازی و تست مدل روی محیط ابری

۵. نتیجه‌گیری و کارهای آتی

در این مقاله مدلی برای پیاده سازی یک الگوریتم مبتنی بر سیستم ایمنی مصنوعی با هدف کشف تقلب کارت های اعتباری ارائه شد. هدف کاهش زمان مورد نیاز سیستم کشف تقلب است چرا که الگوریتم هایی نظیر سیستم ایمنی مصنوعی، علیرغم نتایج بهتری که نسبت به سایر روش ها دارند، زمان آموزش زیادی لازم دارند؛ با توجه به تعداد زیاد تراکنش ها تحمل این زمان در سیستم های حساسی چون کشف تقلب، ممکن نیست. در این مقاله از ابزار Weka برای پیاده سازی مدل در محیط ابری استفاده شده است. با این حال ابزارهای یادگیری ماشینی وجود دارند که در محیط ابری پیاده سازی شده اند. نظیر Mahout که تعدادی از الگوریتم های طبقه بندی و دسته بندی را شامل می شود. به عنوان پیشنهاد می توان الگوریتم سیستم تشخیص ایمنی مصنوعی را به این مجموعه اضافه کرد.

[5] .arff
[6] Skimming
[7] Account Takeover

مشخصه‌های هر گره	CPU	RAM	تعداد گره‌ها	تعداد توابع نگاشت	زمان محاسبه حدآستانه	زمان تولید کشف‌کننده‌ها
بدون موازی‌سازی	2.50 GHz	1GB	۱	-	۱۳۱	۱۶۸
تست ۲	2.50 GHz	512MB	۲	۴	۷۵	۹۱
تست ۳	2.50 GHz	512MB	۲	۱۰	۸۱	۹۴
تست ۴	2.50 GHz	128MB	۴	۴	۸۷	۱۳۳
تست ۵	2.50 GHz	128MB	۴	۱۰	۷۹	۱۳۲

جدول ۱- نتایج پیاده سازی و تست مدل روی محیط ابری

روش	تغییر ایجادشده	نرخ کشف	نرخ FP	نرخ موفقیت	هزینه	هزینه (فرمول ۷)
AIRS	_	0.420	0.021	0.434	479	20827
AIRS	بروزرسانی	0.497	0.018	0.514	177	18092

جدول ۲- نتایج اجرای الگوریتم پایه و الگوریتم بهبودیافته روی مجموعه داده

مراجع

[1] M. Edge, and P. Falcone Sampaio, “A survey of signature based methods for financial fraud detection,” *Computers & Security*, Vol. 28, No. 6, pp. 381-394, 2009.

[2] LN. de Castro, J. Timmis, “Artificial immune systems: A novel paradigm to pattern recognition,” *Artificial Neural networks in pattern Recognition*, University of Paisley, pp. 67–84, 2002.

[3] A. Watkins, J. Timmis, and L. Boggess, “Artificial immune recognition system (AIRS): An immune-inspired supervised machine learning algorithm,” *Genetic Programming and Evolvable Machines*, Vol. 5, No. 3, pp. 291-317, 2004.

[4] M. Gadi, X. Wang, A. Lago, “Credit Card Fraud Detection with Artificial Immune System,” *Artificial Immune Systems Lecture Notes in Computer Science*, Vol. 5132, pp. 119-131, 2008

[5] http://hadoop.apache.org/hdfs/docs/current/hdfs_design.html visited on Feb. 9th 2012.

[6] T. White, *Hadoop: The Definitive Guide*, O’Reilly, 2nd Edition, pp. 15-32, 2011.

[۷] ندا سلطانی، محمدکاظم اکبری، مرتضی سرگلزایی جوان، "متریک جدیدی برای ارزیابی و مقایسه نتایج الگوریتم‌های کشف تقلب در حوزه بانکداری بر اساس هزینه تراکنش"، چهارمین کنفرانس فناوری اطلاعات و دانش، دانشگاه صنعتی انوشیروانی بابل، خرداد ۱۳۹۱