تشخیص رفتار نامتعارف کاربران با استفاده از نگاره رویداد نرم افزار

احسان عسگریان[۱] بهشید بهکمال[۲] محسن کاهانی[۳]

چکیده

هدف این مقاله، ارائه یک فرایند داده کاوی بومی شده برای تشخیص رفتارهای نامتعارف کاربران است. مراحل فرایند پیشنهای براساس یک تجربه عملی است که با استفاده از نگاره رویداد یک سیستم نرم افزاری انجام شده و نتایج نهایی پروژه در این مقاله ارائه شده است. برای اجرای این پروژه، ابتدا شش سناریو برای شناسایی رفتار نامتعارف تعریف شده و نتایج حاصل مورد بررسی قرارگرفته است. سپس با استفاده از روشهای داده کاوی بدون ناظر شامل الگوریتم‌های تحلیل داده‌های غیرمعمول، تحلیل خوشه‌های رفتاری کاربران، تحلیل تغییرات رفتاری کاربران در سیر زمان و تحلیل شباهت عملکرد کاربران، رفتارهای نامتعارف کاربران شناسایی شده است. نتایج پروژه نشان داد که روش ترکیبی بکار گرفته شده برای شناسایی رفتار نامتعارف کاربران از دقت خوبی برخوردار است.

کلمات کلیدی

داده کاوی، تحلیل رفتار کاربر، نگاره رویداد

۱. مقدمه

هدف اصلی داده‌کاوی، دست یافتن به بینش‌هایی است که به لحاظ آماری قابل اعتماد باشند، قبل از انجام داده‌کاوی مجهول بوده و نیز از روی داده‌ها قابل استخراج باشد [1]. به عبارت دیگر داده کاوی فرایند کشف الگوهای مخفی، غیربدیهی، ضمنی، معتبر، جدید و مفید است. در این مقاله، فرایند انجام یک پروژه داده کاوی باهدف شناسایی رفتار نامتعارف کاربران به همراه نتایج حاصل از یک تجربه

[۱] دانشجوی دکتری مهندسی کامپیوتر، نرم افزار، دانشگاه فردوسی مشهد، ehsan.asgarian@mail.um.ac.ir

[۲] استادیار گروه مهندسی کامپیوتر، دانشگاه فردوسی مشهد، behkamal@um.ac.ir (نویسنده مسئول)

[۳] استاد گروه مهندسی کامپیوتر، دانشگاه فردوسی مشهد، kahani@um.ac.ir

عملی ارائه خواهد شد. ساختار مقاله بشرح زیر است: در بخش دوم، مساله مورد بررسی و روش پیشنهادی بیان می‌شود. سپس، اقدامات انجام شده برای شناخت و پیش پردازش داده ها در بخش سوم ارائه می شود. در بخش چهارم، سناریوهای پیشنهادی جهت شناسایی رفتار نامتعارف تعریف شده و نتایج حاصل بطور خلاصه مورد بررسی قرار می‌گیرد. در بخش پنجم، با استفاده از چهار روش داده کاوی، رفتار نامتعارف کاربران شناسایی می‌شود و در بخش ششم مقاله، جمع بندی و نتیجه گیری ارائه خواهد شد.

۲. تعریف مساله و روش پیشنهادی

کلمه‌ی تقلب[۱] به معنی سوء استفاده از منفعت سیستم سازمانی بدون اینکه لزوما به عواقب قانونی مستقیم منجر شود. بطور کلی تشخیص تقلب، بخشی از فرآیند کنترل تقلب می‌باشد که به کاهش فعالیت‌های دستی فرآیندهای نظارت و بررسی کمک کرده و آن‌ها را خودکار می‌کند. البته تقلب در هر کسب‌وکار با توجه به قوانین و آیین‌نامه‌های موجود معنی پیدا می‌کند. برای مثال ورود به بانک ساعت ۱ بعد از ظهر امری بسیار طبیعی است درحالیکه ورود به بانک ساعت ۱ صبح (نصف شب) ممکن است تقلب (تخلف) محسوب شود. ولی این قاعده درباره بیمارستان صادق نیست. هدف این مقاله، شناسایی موارد غیرعادی، یا رفتارهای متفاوت با دیگران است. برای این منظور از دو روش استفاده شده است: ۱) تعریف سناریو برای تحلیل رفتار نامتعارف کاربران، و ۲) استفاده از روشهای داده کاوی برای تشخیص رفتار نامتعارف. در بخش بعدی، سناریوهای پیشنهادی تعریف شده و نتایج حاصل ارائه می‌شود.

۳. شناخت داده‌ها و پیش‌پردازش فایل رویداد

مجموعه داده مورد بررسی، یک فایل نگاره رویداد است که توسط نرم افزار بصورت خودکار تولید شده است. حجم اطلاعات بعد از پیش‌پردازش اولیه حدود ۲,۵ گیگابایت بوده است. در این فایل رویداد، ۱۲ ویژگی موجود است و تعداد رکوردهای فایل، حدود ۱۱ میلیون رکورد می باشد. بنابراین حجم داده برای انجام پروژه مناسب است. فهرست ویژگی‌های داده‌های فایل رویداد بهمراه اطلاعات اولیه درباره آنها در جدول (۱) آمده است:

---

[۱] Fraud

جدول (۱) ویژگی‌های موجود در فایل رویداد سیستم

نام ستون	توضیح	نوع داده	نمونه داده	تعداد مقادیر یکتا	تعداد مقادیر تهی (NULL)
BizCodeTitle	عنوان	رشته	شماره ارجاع	26	142,772
BizCode	کد رهگیری	رشته	900113425	1,360,873	875,854
BizCodeType	نوع شماره	رشته	NidKartabl	915	1,332,514
NosaziCode	کد نوسازی	رشته	9-11-105-14-1-0-0	972,432	2,656,754
NidKartabl	شماره پرونده	عدد	900113425	756,367	5,505,083
UserName	نام کاربر	رشته	محسن داوودی - davoodi	56,904	346,604
LogDate	تاریخ عملیات	رشته	1391/02/09	1,481	0
LogTime	زمان عملیات	رشته	0.692361111	1,440	0
IP	ای پی	رشته	192.168.90.172	173,418	0
FormCaption	نام فرم برنامه	رشته	تایید مدیران	312	119,527
WorkFlowTyp e	نوع گردش کار	رشته	* گواهی پایان کار بهره برداری	211	5,488,351
Action	ضمیمه XML (تغییرات فیلدها)	عدد	فایل XML که تغییرات فیلدهای فرم را نشان میدهد	؟	0

اقدامات انجام گرفته برای پیش‌پردازش داده‌ها به سه دسته زیر تقسیم میشوند:

- **تبدیل ویژگی‌ها:** در اولین مرحله پیش پردازش، ابتدا تعداد تغییرات فیلدهای فرم، بجای مقدار XML جایگزین شده است. سپس رشته تاریخ و زمان به عدد اعشاری تبدیل شده است.
- **استخراج ویژگی‌های جدید:** در این مرحله، برای درک و تفسیر بهتر نتایج، ویژگیهای جدید استخراج شده است. مانند استخراج شماره ماه ۱ الی ۵۰، استخراج شماره روز ۱ الی ۱۴۸۱، استخراج محدوده IP با استفاده از دو بخش ابتدایی آدرس IP، افزودن فیلد رده کاربر براساس تعداد لاگ موجود در فایل رویداد و افزودن فیلد رده تغییرات فیلدهای فرم.
- **پاکسازی داده ها:** در آخرین گام پیش پردازش، ابتدا ویژگی هایی که خاصیت کلید داشته و کارکرد غیرتوصیفی دارند، حذف شده اند. سپس لاگ مربوط به کاربرانی که

کمتر از ۱۰ لاگ در فایل رویداد دارند، داده پرت شناخته شده و از فایل رویداد حذف شده است.

بعد از انجام پیش‌پردازش، فاز تحلیل آماری داده‌ها به منظور شناخت بهتر داده‌ها و کشف موارد غیر متعارف در داده‌ها لاگ آغاز شد.

۴. تعریف سناریو برای تشخیص رفتار نامتعارف و تحلیل نتایج

همانطور که اشاره شد، هدف این بخش شناسایی بهتر داده‌ها و کشف روابط ساده و الگوهای غیرمتعارف اولیه در بین داده‌ها با توجه به کسب‌وکار سازمان است. برای این منظور، ابتدا ویژگی‌های مهم شامل نام کاربر، IP و عنوان فرم مشخص شده‌اند. سپس سناریوهای موارد غیرعادی براساس ویژگی‌های محوری تعیین شد. سناریوهای تعریف شده و نتایج حاصل بصورت خلاصه در زیر آورده شده است:

۴.۱.شناسایی کاربرانی که از مکانهای جغرافیایی مختلف با سیستم کار می‌کردند

با توجه به اینکه سیستم مورد مطالعه توسط واحدهای سازمانی که در مکانهای جغرافیایی مختلف قرار دارند، قابل دسترسی است، در این سناریو کاربرانی که از مکانهای مختلف به سیستم متصل شده اند، شناسایی می‌شوند. نتایج نشان داد که تعداد کمی از کاربران با بیش از ۲۰ محدوده IP مختلف (از بیش از ۲۰ مکان جغرافیایی مختلف) به سیستم متصل شده اند؛ در حالیکه اغلب کاربران (بیش از ۹۰٪ کاربران) تنها با یک یا دو محدوده IP مختلف در سیستم فعالیت می‌کنند.

۴.۲.تحلیل رفتار کاربران براساس تاریخ و زمان استفاده از سیستم

براساس نتایج بدست آمده، مشخص شد که برخی کاربران در ساعت نیمه شب در سیستم فعالیت داشتند که مجموعاً حدود ۱۰۲ داده لاگ تولید کردند. همچنین مشخص شد که این کاربران بیش از ۹۰٪ فعالیت‌های خود را شبانه (بین ۱۱ شب تا ۵ صبح) انجام می‌دهند.

۴.۳. کاربرانی که بیشترین لاگ مربوط به روزهای تعطیل (روزهایی با کمترین لاگ) داشتند

در اجرای این سناریو، ابتدا روزهایی که کمترین تعداد لاگ در سیستم را داشتند به عنوان روزهای تعطیل (کم کار) مشخص شد. سپس تحلیلی برای کاربرانی که اغلب در این روزها بیشترین فعالیت را

داشتند، انجام شد. در این تحلیل مشخص شد که برخی کاربران به اندازه قابل توجهی نسبت به سایرین در روزهای تعطیل با سیستم کار می‌کنند.

۴.۴.کاربرانی که اغلب در یک واحد سازمانی خاص کار می‌کردند ولی گاهاً پرونده های سایر واحدهای سازمانی را نیز تغییر دادند.

در ادامه تحلیل فعالیت کاربران در مناطق مختلف، در این قسمت تحلیلی بر روی کاربرانی که در رفتار آنها از نظر فعالیت بر روی ادارات مختلف نامتوازن بوده، انجام شده است. هدف این تحلیل نشان دادن فعالیت‌های کاربرانی است که اغلب بر روی یک یا دو واحدسازمانی خاص فعالیت داشتند ولی گاهاً برروی تعداد اندکی از پرونده‌های سایر واحدها نیز فعالیت داشتند. این رفتار متفاوت با رفتار اکثریت کاربران و رفتاری خارج از عرف شناسایی شده است.

۴.۵. فرم‌هایی که توسط تعداد کمی از کاربران، دسترسی قابل توجهی داشته اند.

در این قسمت تحلیلی بر روی فرم‌های خاص انجام شده است. منظور از خاص بودن یک فرم این است که تعداد قابل توجهی لاگ برای یک فرم توسط تعداد محدودی از کاربران ایجاد شده باشد. نتایج نشان داد که برای یک فرم خاص که کمتر از ۱۰ کاربر به آن دسترسی داشته اند، حدود ۱۱.۰۰۰ خط لاگ در فایل رویداد موجود است.

۴.۶.تحلیل ورود ناموفق به سیستم

در این سناریو کاربرانی که بیشترین ورود ناموفق به سیستم را داشتند، به عنوان رفتاری خارج از عرف تشخیص داده شده است.

۵. استفاده از روش‌های داده کاوی برای تحلیل رفتار نامتعارف

در این مرحله، روش‌های داده کاوی مناسب انتخاب و اجرا می‌شوند. به منظور تشخیص موارد غیرعادی، چهار رویکرد اصلی وجود دارد [2]: رویکرد با ناظر،[۱] رویکرد بدون ناظر،[۲] رویکرد ترکیبی روی داده‌های برچسب‌گذاری شده و رویکرد نیمه نظارتی[۳] روی داده‌های قانونی (بدون تقلب). در این پروژه،

[۱] Supervised

[۲] Unsupervised

[۳] Semi-Supervised

با توجه به عدم وجود داده‌های نمونه آموزشی (معرفی نمونه‌های داده‌ای یا رفتاری غیرمتعارف)، از رویکرد بدون ناظر استفاده شده است. روشهای تشخیص رفتارهای نامتعارف با رویکرد بدون ناظر عبارتند از:

- الگوریتم‌های استاندارد تحلیل داده‌های غیرمعمول
- تحلیل گروه‌های (خوشه‌های) رفتاری کاربران
- تحلیل تغییرات رفتاری کاربران در سیر زمان
- تحلیل شباهت عملکرد کاربران و شناسایی موارد غیرمعمول

به منظور اجرای روشهای فوق در این پروژه از ابزار IBM SPSS Modeler (SPSS Clementine) استفاده شده است. IBM SPSS Modeler، نرم افزاری از شرکت SPSS است که در ابتدا با نام کلمنتاین (Clementine) ارائه می‌شد که بعد از نسخه ۱۳ به SPSS Modeler تغییر نام پیدا کرده است. این نرم‌افزار، یکی از بهترین ابزارهای داده کاوی است و برنامه‌ای حرفه ای برای انجام محاسبات پیچیده و آنالیزهای آماری است. در ادامه نتایج هریک از چهار روش پیشنهادی ارائه خواهد شد.

۵,۱. الگوریتم‌های استاندارد تحلیل داده‌های غیرمعمول

یکی از روش‌های متداول شناسایی داده‌های خارج از عرف (پرت)، خوشه‌بندی داده‌ها و شناسایی داده‌های متفاوت به ازای هر خوشه هستند. به عبارت دیگر، داده‌هایی که از مراکز خوشه‌ها بسیار دور باشند، نشان دهنده داده‌هایی هستند که نسبت به سایرین رفتار متفاوتی دارند. در این بخش چند الگوریتم پرکاربرد برای شناسایی داده‌های پرت معرفی می‌شوند؛ سپس، الگوریتم مورد استفاده برای شناسایی داده‌های پرت به همراه مراحل پیاده‌سازی این رویکرد در ابزار IBM SPSS Modeler و تحلیل نتایج حاصل، ارائه خواهند شد. از جمله الگوریتم‌هایی که برای شناسایی داده‌های پرت مورد استفاده قرار می گیرند، می توان به الگوریتم‌های زیر اشاره نمود:

- **Local Outlier Factor (LOF)** [3]: این الگوریتم که توسط Markus M. Breunig و همکارانش در سال ۲۰۰۰ معرفی شد، داده‌های پرت را از طریق اندازه‌گیری انحراف محلی هریک از داده‌ها با داده‌هایی که در همسایگی آن هستند، پیدا می‌کند. این الگوریتم همانطور که از عنوان آن بر می‌آید، بر اساس مفهوم چگالی محلی عمل می کند. محلی بودن با KNN بدست می آید و فاصله بدست آمده برابر با چگالی تخمین زده شده می‌باشد. از طریق

مقایسه‌ی چگالی محلی یک داده با چگالی محلی همسایه های آن، منطقه‌ی داده‌هایی که چگالی مشابه دارند شناسایی می‌شود، و داده‌هایی که نسبت همسایه های خود در این منطقه، بصورت قابل توجهی چگالی پایین تری داشته باشند، بعنوان داده‌های پرت شناسایی می‌شوند. چگالی محلی به کمک نوعی اندازه‌گیری فاصله هر نقطه تا نقطه‌های همسایه آن محاسبه می‌شود.

- **K-NN Global Anomaly Score** [4]: در این الگوریتم داده‌های پرت براساس فاصله بین هر داده تا k-امین داده در همسایگی آن، شناسایی می‌شوند. درواقع، یک مقدار امتیاز (score) براساس میانگین فاصله‌ی k همسایه‌ی یک داده محاسبه می‌شود که به کمک آن داده‌های پرت شناسایی می‌شوند.
- **Connectivity-based Outlier Factor (COF)** [5]: این الگوریتم یک درجه به هر داده (نقطه) می‌دهد که به آن، عامل پرت مبتنی بر اتصال گفته می‌شود. داده‌هایی که میانگین اتصال محلی کم‌تری نسبت به همسایه‌های خود دارند، بعنوان داده‌های پرت شناسایی می‌شوند. داده‌هایی که متعلق به یک کلاس هستند، معمولا مقدار COF کمتری دارند.
- **Local Correlation Integral (LOCI)** [6]: الگوریتم LOCI یا انتگرال همبستگی محلی، یک الگوریتم قوی در شناسایی داده‌های پرت می‌باشد که سرعت بالایی در پردازش داده‌های با حجم بالا دارد. این الگوریتم، ابعاد داده‌های اطراف هر داده را اندازه‌گیری می‌کند (تعداد همسایه‌ها). داده‌های پرت در این الگوریتم براساس میزان تفاوت قابل توجهی که با همسایه‌های خود دارند، شناسایی می‌شوند. این روش، نه تنها داده‌های پرت بلکه خوشه‌های کوچک را نیز شناسایی می‌کند.
- **Local Outlier Probability (LoOP)** [7]: مانند LOF این الگوریتم به داده‌ها یک مقدار بین صفر و یک می‌دهد با این تفاوت که می‌توان این مقدار را بصورت یک مقدار احتمالی برای شناسایی داده‌های پرت و احتمال پرت بودن این داده‌ها تفسیر کرد. این الگوریتم، در سال ۲۰۰۹ و توسط Kriegel و همکارانش ارائه گردیده است.
- **Cluster-based Local Outlier Factor (CBLOF)** [8]: در این الگوریتم به هر داده یک عدد عامل پرت با نام (CBLOF) تخصیص داده می‌شود که با استفاده از دو مقدار

خوشه‌ای که داده به آن تعلق دارد و فاصله بین هر داده تا نزدیک‌ترین خوشه به آن محاسبه می‌شود. برای محاسبه فاصله بین داده و خوشه در این الگوریتم، باید معیار شباهت در الگوریتم خوشه‌بندی مورد استفاده قرار گرفته شده، تنظیم شود.

برای پیاده‌سازی این سناریو از الگوریتم شناسایی ناهنجاری[1] در نرم افزار IBM SPSS Modeler استفاده شده است. الگوریتم‌های شناسایی ناهنجاری، در شناسایی رفتار غیر معمول مانند مواردکلاهبرداری یا شناسایی موارد مشکوک کاربرد دارند. در این الگوریتم‌ها، داده‌هایی که از مراکز خوشه‌ها بسیار دور باشند، نشان دهنده داده‌هایی هستند که نسبت به سایرین رفتار متفاوتی دارند. بعد از اجرای الگوریتم، مهم‌ترین ویژگی‌ها برای تشخیص رفتارهای نامتعارف به ترتیب زیر شناسایی شد: ۱) واحد سازمانی، ۲) محدود تاریخ یا شماره ماه، ۳) محدوده IP، ۴) تعداد تغییرات در فرم و ۵) زمان رخداد فعالیت در شبانه‌روز.

۵,۲. تحلیل خوشه‌های رفتاری کاربران

با استفاده از خوشه‌بندی داده‌های فایل رویداد، داده‌هایی که مشابه هم هستند در خوشه‌ای یکسان قرار می‌گیرند. پس هدف از خوشه‌بندی به نوعی دسته‌بندی رفتاری داده‌های لاگ (فعالیت کاربران) است. در این رویکرد هدف شناسایی کاربرانی است که رفتار مشخصی ندارند. به عبارت دیگر هدف شناسایی کاربرانی است که لاگ فعالیت آنها در سیستم در خوشه‌های رفتاری مختلف قرار می‌گیرند. بدین منظور مراحل زیر به ترتیب انجام شدند:

۱. اجرای روش‌های خوشه‌بندی متفاوت (با پارامترهای مختلف) بر روی داده‌های فایل رویداد
۲. شناسایی برچسب خوشه‌های مختلف کاربران
۳. استخراج کاربرانی که در بیشترین تعداد خوشه‌ها حضور داشتند به عنوان کاربران غیرمعمول
۴. استخراج کاربران مشترک در نتایج کاربران غیرمعمول ایجاد شده توسط روش‌های خوشه‌بندی مختلف
۵. استفاده از درخت تصمیم برای تحلیل و توصیف ویژگی‌های کاربران با رفتار غیرمتعارف

[1] Anomaly detection

پس از انجام مراحل فوق، مجموعه از قوانین برای تشخیص رفتار نامتعارف استخراج شد. به عنوان مثال یک قانون استخراج شده بصورت زیر است: "رفتار کاربری که در بیش از دو «عنوان پرونده» فعالیت داشته و کمتر از ۲۱ «گردش کار» مختلف فعالیت داشته و با بیش از ۱۱۲ «IP» مختلف به سیستم وصل شده، نامتعارف هست." به بیان دیگر "کاربری که روی پرونده‌های کمی کار کرده ولی با تعداد IP مختلف زیادی (از مکانهای متفاوت) به سیستم وصل شده، رفتاری غیرمتعارف دارد."

۵,۳. تحلیل تغییرات رفتاری کاربران در سیر زمان

هدف از اجرای این روش، شناسایی کاربرانی است که در طول زمان رفتاری متغیر دارند. مشابه روش قبل، برای شناسایی الگوهای رفتاری کاربران، عمل خوشه‌بندی بر روی داده‌های فایل رویداد انجام شد. سپس داده‌ها بر حسب زمان مرتب و به ۱۰ بازه زمانی برابر تقسیم شدند. در داده‌های فایل رویداد متعلق به هر کاربر در هر بازه، بیشترین برچسب خوشه متعلق به هر کاربر به عنوان گروه رفتاری آن کاربر در آن بازه زمانی تعیین شد. بدین منظور مراحل زیر به ترتیب انجام شدند:

۱. اجرای روش خوشه‌بندی متفاوت بر روی داده‌های رویداد
۲. مرتب‌سازی داده‌های لاگ براساس زمان و تقسیم داده‌های لاگ به ۱۰ بازه زمانی مختلف
۳. انتساب یک برچسب خوشه رفتاری (با بیشترین تکرار) برای هر کاربر در هر بازه زمانی
۴. استخراج کاربرانی که در سیر زمان (بازه‌های زمانی مختلف) بیشترین تغییر رفتاری را داشتند به عنوان کاربران غیرمعمول
۵. استخراج کاربران مشترک در نتایج روش‌های خوشه‌بندی مراحل قبل
۶. تحلیل و توصیف ویژگی‌های کاربرانی که با رفتار غیرمتعارف شناسایی شدند، با استفاده از درخت تصمیم

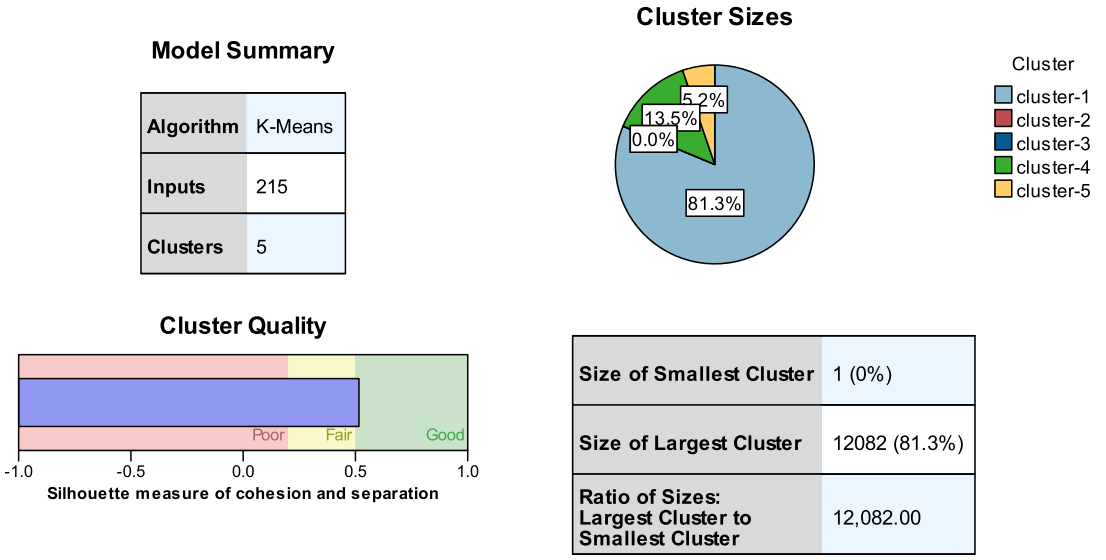
برای تحلیل نتایج و شناسایی سلسله عوامل و قوانین حاکم بر روی نتایج (عوامل تعیین کننده رفتارهای خارج از عرف کاربران)، مشابه رویکرد قبل از روش‌های درخت تصمیم استفاده شد و در پایان ۹ کاربر بعنوان کاربران با رفتار غیرمتعارف شناسایی شدند.

۵,۴. تحلیل شباهت عملکرد کاربران و شناسایی موارد غیرمعمول

در این روش، تحلیلی بر عملکرد کاربران انجام شد. سپس خوشه‌بندی روی ویژگی‌های عملکردی (پروفایل عملکرد) کاربران انجام شد. هدف از این خوشه‌بندی تعیین گروه‌های کاربرانی است که رفتارهای

مشابه دارند. در نهایت کاربرانی که رفتار آنها فاصله قابل توجهی از مراکز خوشه‌های رفتاری سایر کاربران داشته باشد به عنوان کاربران با رفتار خارج از عرف شناسایی می‌شود. اولین گام برای انجام این تحلیل ایجاد جدول پروفایل عملکردی کاربران است. به ازای هر یک از این کاربران یک سطر در جدول پروفایل عملکردی کاربران در نظر گرفته شد. ویژگی‌هایی که برای هر کاربر از روی داده‌های لاگ جمع‌آوری کردیم ترکیبی از اطلاعات فیلدهای داده‌های لاگ (تعداد مقادیر غیر تکراری) بیشترین مقدار تکراری هر فیلد، گروه‌بندی‌های انجام شده برای کاربران و اطلاعات فرم‌های بازدید شده در نرم افزار هستند.

براساس نتایج حاصل از خوشه بندی، داده‌ها در ۵ خوشه گروه‌بندی شده‌اند. خوشه ۱ شامل ۸۱٪ داده‌هاست، یعنی اغلب کاربران (۸۱٪) رفتاری مشابه به هم دارند. نکته جالبی که در خروجی این رویکرد دیده می‌شود مربوط به خوشه‌های ۲ و ۳ است. در هر یک از خوشه‌های ۲ و ۳ تنها یک داده (کاربر) وجود دارد. این امر بدین معنی است که دو کاربر هستند که رفتاری متفاوت نسبت به سایر کاربران دارند. خوشه‌های ۴ و ۵ نیز به ترتیب شامل ۱۳,۵٪ و ۵,۲٪ از کاربران هستند. نتایج این خوشه بتدی در شکل (۱) آمده است.



شکل (۱) اطلاعات خوشه‌بندی اطلاعات پروفایل عملکردی کاربران

۶. نتیجه گیری

در این مقاله، فرایند اجرا و نتایج حاصل از یک پروژه عملی داده کاوی با هدف تشخیص رفتار نامتعارف کاربران ارائه شد. ابتدا با بیان مساله مورد بررسی، روش پیشنهادی توضیح داده شد.

سپس، اقدامات انجام شده برای پیش پردازش داده ها، شامل تبدیل ویژگی ها، استخراج ویژگی جدید و پاکسازی داده ها، ارائه شد. پس از آن، شش سناریو پیشنهادی برای شناسایی رفتار نامتعارف تعریف شده و نتایج حاصل مورد بررسی قرارگرفت. همچنین با استفاده از روشهای داده کاوی بدون ناظر شامل الگوریتم‌های تحلیل داده‌های غیرمعمول، تحلیل گروه‌های رفتاری کاربران، تحلیل تغییرات رفتاری کاربران در سیر زمان و تحلیل شباهت عملکرد کاربران، رفتار کاربران تحلیل و موارد نامتعارف استخراج شد.

۷. مراجع

[1] C. Elkan, "Magical Thinking in Data Mining: Lessons from CoIL Challenge 2000," in *Proc. of SIGKDD01*, 2001.

[2] C. Phua, V. Lee, K. Smith and R. Gayler, "A comprehensive survey of data mining-based fraud detection research," in *arXiv preprint*, 2010.

[3] "https://en.wikipedia.org/wiki/Local_outlier_factor," 2015. [Online].

[4] M. Amer and M. Goldstein, "Nearest-Neighbor and Clustering based Anomaly Detection Algorithms for RapidMiner," in *Proceedings of the 3rd RapidMiner Community Meeting and Conferernce (RCOMM 2012)*, 2010.

[5] D. Pokrajac, N. Reljin, N. Pejcic and A. Lazarevic, "Incremental Connectivity-Based Outlier Factor Algorithm," in *BCS International Aca demic Conference 2008 – Visions of Computer Science*, 2008.

[6] S. Papadimitriou, H. Kitagawa, P. B. Gibbons and C. Faloutsos, "Loci: Fast outlier detection using the local correlation integral," Tech. Rep. IRP-TR-02-09, Intel Research Laboratory, 2002.

[7] H.-p. Kriegel, P. Kröger, E. Schubert and A. Zimek, "LoOP: Local Outlier Probabilities," in *ACM CIKM'09*, Hong Kong, China, 2009.

[8] Z. He, X. Xu and S. Deng, "Discovering Cluster Based Local Outliers," *Pattern Recognition Letters*, vol. 24, no. 9, pp. 1641-1650, 2003.

[9] N. Lavrac, H. Motoda, T. Fawcett, R. Holte, P. Langley and P. Adriaans, "Introduction: Lessons Learned from Data Mining Applications and Collaborative Problem Solving," *Machine Learning*, vol. 57, pp. 13-34, 2004.