مطالعات آینده پژوهی و سیاستگذاری

دوره ۳، شماره ۲، تابستان ۱۳۹۶

صفحات ۷۷- ۶۵

Online ISSN: ٢٤٧٦-٤٤٤٢

Print ISSN: ٢٥٣٨-٢٠٤٧

www.irijournals.com

بررسی و شناسایی الگوریتم پیشنهادی در زمینه جرائم الکترونیکی در وب سایت پورتال دیوان محاسبات کشور

مینا جانتنی

کارشناس نرم افزار کامپیوتر، دیوان محاسبات استان کرمان، کرمان، ایران

چکیده

جرایم الکترونیکی و مجازی قدمت کوتاهی دارند و تنها طی بیست سال اخیر این اصطلاح رواج یافته است و با ساده تر شدن کاربرد و استفاده از رایانهای برای همگان و کاهش قیمت دسترسی به ابزار فناوری اطلاعات، معظلی نوین به نام جرایم الکترونیکی در فضای مجازی پدید آمده است و نهادهای امنیتی، انتظامی و نظارتی را با چالش جدیدی مواجه ساخته است. جرایم الکترونیکی که به جرایم نسل سوم رایانه و اینترنت وابسته است، در محیط مجازی یا فضای الکترونیکی قابل تحقق می‌باشد. در این میان راهکارهای علمی و هوشمند کشف جرم، به دلیل برخورداری از پشتوانه علمی و دانش ریاضی توجه بسیاری از جرم شناسان را به خود معطوف داشته است. یکی از این راهکار ها، داده کاوی می‌باشد. داده کاوی فرایندی است که با استفاده از روشهای هوشمند، دانش را از مجموعه ای از داده ها استخراج می کند. دراین مقاله با استفاده از روش تحلیل محتوا و بررسی اسناد با بکارگیری روش ماشین بردار پشتیبان سیستم هوشمندی طراحی گردید تا جرایم الکترونیکی را در یک محیط مجازی تشخیص دهد. الگوریتم پیشنهادی را برای شناسایی رفتار کاربران وب سایت پورتال دیوان محاسبات کشور و تشخیص رفتارهای مشکوک با شدتهای مختلف را تست نمودیم. بدین منظور کلیهٔ رفتارهای کاربران را با استفاده از داده کاوی و روش تحلیل محتوا در پنج سطح مختلف دسته بندی نمودیم. سپس سیستم بردار پشتیبان را براساس این کلاس ها آموزش و تست نمودیم.

واژه‌های کلیدی: داده کاوی، جرایم الکترونیکی، جرایم الکترونیکی، جرم، دیوان محاسبات کشور.

۱. مقدمه

گستره وسیع کاربردهای داده کاوی این موضوع را به عنوان یکی از مهم ترین حوزه های پژوهشی مطرح نموده است. یکی از عرصه های مهم کاربرد داده کاوی، حوزه جرم شناسی است. جرم شناسی به فرآیند تحلیل و آنالیز یک جرم و شناخت ویژگیهای آن اطلاق میشود. (برون[1]، ۲۰۰۸) و اما جرایم الکترونیکی و مجازی قدمت کوتاهی دارند و تنها طی بیست سال اخیر این اصطلاح رواج یافته است و با ساده تر شدن کاربرد و استفاده از رایانه ای برای همگان و کاهش قیمت دسترسی به ابزار فناوری اطلاعات، معظلی نوین به نام جرایم الکترونیکی در فضای مجازی پدید آمده است و نهادهای امنیتی، انتظامی و نظارتی را با چالش جدیدی مواجه ساخته است. جرایم الکترونیکی که به جرایم نسل سوم رایانه و اینترنت وابسته است، در محیط مجازی یا فضای الکترونیکی قابل تحقق می باشد. از این جهت رفتار مجرمان رایانه ای کاملاً متفاوت از مجرمان سنتی است. در این نوع از بزه، مرتکبان ناشناس در فضایی ناشناخته دست به اعمال مجرمانه می زنند. در شیوه جدید و جرایم نوین رایانه ای از بزهدیده به صورت ماشین - ماشین تغییر یافته است که بیشترین مورد تحقق آن در جرایم تجارت الکترونیکی و جرایم الکترونیکی است در این نوع از بزه، مرتکبان ناشناس در فضایی ناشناخته دست به اعمال مجرمانه می زنند و همچنین مجرمین رایانه ای و مجازی فارغ از زمان و مکان بوده و این نوع از جرایم و تخلفات هماکنون جنبه فراملی و فرا سرزمینی به خود گرفته است. فناوریهای نوین در این عرصه و پیشرفت تجهیزات ارتباطی، مخابراتی و الکترونیکی، سهولت وقوع جرایم رایانه‌ای در فضای مجازی، متخلفان و مجرمان را قادر ساخته که فعالیتهای خود را بدون داشتن ارتباطی خاص با یک محل معین و مشخص انجام دهند و همچنین استفاده از سیستم های الکترونیکی در موسسات و سازمان های دولتی و خصوصی جهان به سرعت رو به گسترش بوده و شمار استفاده کنندگان از خدمات الکترونیکی روز به روز در حال افزایش است از این رو سازمان ها تلاش دارند تا با بکارگیری اینترنت به عنوان یکی از کانالهای اصلی ارائۀ خدمات، برای نفوذ و قدرتمند ساختن کسب و کار خود استفاده کند خدمات اینترنتی در دو حالت به انجام فعالیتهای مرتبط با کسب و کار می پردازد.

۲- بیان مسأله

در جامعه الکترونیکی امروز، تجارت الکترونیک تبدیل به یکی از کانال های مهم و اصلی در زمینه ی فروش در تجارت جهانی شده است. تاکنون تلاش زیادی توسط محققان در رشته های مختلف برای کشف دانش مخفی در مجموعه داده‌ها صورت گرفته است. همان گونه که مشاهده می شود داده کاوی یکی از گام های این فرآیند است. بر اساس مدل CRISP مراحل اصلی در فرایند داده کاوی جهت کشف دانش عبارتند از: (ژیاوی هان،۱۳۹۲)

درک کسب و کار: پییش زمینه داده کاوی، داشتن درک صحیحی از داده ها و تعریف مسأله است.

درک داده ها: در این مرحله باید، داده های اولیه جمع آوری شده مورد تفسیر و کاوش قرار گیرند

آماده سازی داده برای مدل سازی: پس از گذراندن دو مرحله پیشین، باید داده های جمع آوری شده پیش پردازش شوند.

ساختن مدل: در قالب این گام باید مراحلی از قبیل انتخاب تکنیک و روش مدل سازی، ساخت مدل و بررسی و سنجش مدل انجام شوند.

ارزیابی مدل: پس از اینکه مدل مربوطه ساخته شد، باید نتایج آن مورد ارزیابی قرار گیرند.

استفاده علمی از دانش کشف شده: در قالب این گام باید در خصوص مسائلی از قبیل، چگونگی استفاده از نتایج حاصل از داده کاوی، تشخیص افرادی که نیاز به استفاده از نتایج دارند و تعداد دفعاتی که نیاز به استفاده از نتایج است، مورد تصمیم گیری قرار گیرند. (حاجیان، ۲۰۱۱)

[1] Brown

۳- جرایم رایانه ای

جرم های رایانه ای و الکترونیکی خود دارای انواعی می باشد که بسته به ماهیت آن ها دسته بندی شده اند. در زیر به برخی از انواع آن پرداخته شده است:

هک کردن: عبارت است از نفوذ به یک سیستم کامپیوتری بدون داشتن مجوز.

فیشینگ[1]: فیشینگ عبارت است از تلاش برای بدست آوردن اطلاعاتی مانند رمز عبور، شناسه عبور و جزئیات کارت اعتباری با جا زدن خود به عنوان یک منبع قابل اعتماد.

هویت جعلی: هویت جعلی یا خود را به جای کس دیگر شناساندن یکی از جدیدترین کلاه برداری هایی است که به کمک آن پول های زیادی ربوده شده و سودهای کلانی نصیب کلاه برداران می شود.

۳-۱ داده کاوی و جرایم الکترونیکی

ماهیت پیچیده داده های مرتبط با جرم و بزهکاری و روابط نامحسوس میان این داده ها موجب مقبولیت روز افزون استفاده از دانش داده کاوی در میان جرم شناسان و تحلیلگران جرم شده است. در واقع دانش حاصل از اعمال روش های داده کاوی در حوزه تحلیل جرم بستر مناسبی را برای پشتیبانی اطلاعاتی فرماندهان و مدیران به منظور انجام به فعالیت های آتی پلیس فراهم می آورد.

۳-۲- محدودیت های داده کاوی

اولین مورد به سخت افزار و نرم افزار لازم و موقعیت وب سایت پورتال دیوان محاسبات کشور اطلاعاتی مربوط می شود. برای مثال درکشور هند، داده های غیر مجتمع که برای کاربردهای داده کاوی لازم است ممکن است به فرم دیجیتالی در دسترس نباشد. (کاظمی، ۱۳۸۸)

۴- حفاظت از حریم شخصی در سیستم های داده کاوی

داده کاوی با استخراج موفقیت آمیز اطلاعات، دانش مورد نیاز برای استفاده در زمینه های مختلف از جمله بازاریابی، هواشناسی، تحلیل های پزشکی و امنیت ملی را فراهم می سازد؛ ولی هنوز هیچ تضمینی ارایه نشده است که بتوان داده‌های خاصی را مورد داده کاوی قرار داد؛ بدون آن که به حریم خصوصی مالک آن اطلاعات تجاوز کرد.

۵- اهمیت و ضرورت تحقیق

در جهان تجارت و اقتصاد، اطلاعات مربوط به مشتریان و شرکت های حقوقی، دارایی های استراتژیک محسوب می‌شوند، از این رو در دنیای رقابتی امروز استخراج دانش از این اطلاعات و تصمیم گیری بر اساس آنها از اهمیت بسیار بالایی برخوردار است.

۶- تعاریف و مفاهیم

۶-۱ داده:

داده می تواند هر نوع اطلاعاتی باشد. این اطلاعات می تواند محتویات یک سلول بر روی یک صفحه ی گسترده، یک فایل صوتی یا ویدئوئی، یک رشته از کلمات در یک سند، اطلاعات خام ایجاد شده بعنوان خروجی یک برنامه کامپیوتری و یا اطلاعات مورد استفاده برای توصیف یک فایل باشد.

[1] Phishing

۶-۲ تجمیع داده

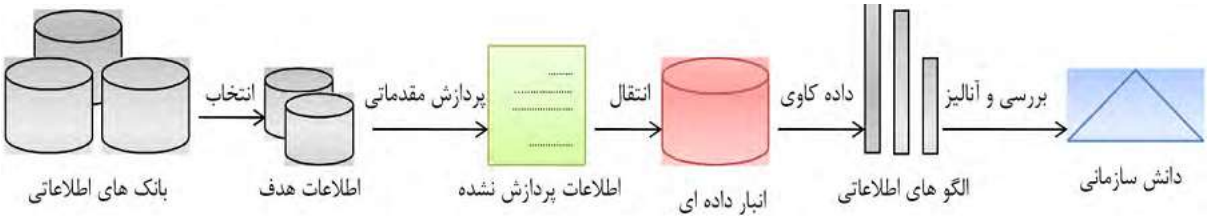
روش های مختلفی برای دسته بندی وب سایت پورتال دیوان محاسبات کشور های اطلاعاتی وجود دارد. بعضی از مردم ترجیح می دهند که آنها را براساس نوع داده هایی که در وب سایت پورتال دیوان محاسبات کشور های اطلاعاتی ذخیره می شوند، طبقه بندی کنند.

۶-۳ پایگاه داده

یک پایگاه داده عملاً یک وب سایت پورتال دیوان محاسبات کشور اطلاعاتی است که اطلاعات سایر وب سایت پورتال دیوان محاسبات کشور های اطلاعاتی را با استفاده از یک فرمت مشترک ذخیره سازی می نماید.

۶-۴- داده کاوی

داده کاوی فرایند کشف ارتباطات، طرح ها و الگوهای جدید معنی دار در پایگاه های اطلاعاتی می باشد



شکل ۱: فرایند داده کاوی

۶-۵ قوانین تلازمی در داده کاوی

کشف دانش در پایگاه داده فرایند شناسایی درست، ساده، مفید و در نهایت الگوها و مدل های قابل فهم در داده ها است که داده کاوی مرحله ای از فرایند کشف دانش است و شامل الگوریتم های مخصوص به خود است، به طوری که، تحت محدودیت های مؤثر محاسباتی قابل قبول، الگوها و مدل ها را در داده ها کشف می کند.

۶-۶ تکنیک های داده کاوی

دسته بندی

متداول ترین تکنیک است و یکسری نمونه های از پیش تعیین شده را شامل می شود که برای توسعه مدل به کار می رود که بتواند انواعی از موارد ثبت شده را دسته بندی نماید.

خوشه بندی

داده ها ممکن است حاوی ساختارهای پیچیده ای باشند که حتی بهترین تکنیکهای دادهکاوی هم قادر به استخراج الگوهای معنیدار از آنها نباشند (هان[1]، ۲۰۱۳).

رگرسیون گیری

تکنیک رگرسیون گیری را می توان برای پیش بینی پذیرفت. از تحلیل رگرسیون می توان برای مدل سازی روابط یک یا چند متغیر مستقل و وابسته استفاده نمود.

تجمع و همبستگی

تجمع و همبستگی معمولاً برای یافتن یک آیتم تکراری به کار می رود و یافته ها را در میان مجموعه اطلاعات و سیستم تنظیم می کند.

[1] Han

درخت تصمیم گیری

درخت تصمیم درختی است که در آن نمونه ها را به نحوی دسته بندی می کند که از ریشه به سمت پائین رشد می کنند و در نهایت به گره های برگ می رسد.

الگوریتم ژنتیک

الگوریتم ژنتیک (GA) یک جستجوی اکتشافی است که فرآیند تکامل طبیعی را پیروی میکند، این اکتشاف برای ایجاد راه حل های مفید در بهینه سازی و جستجوی مسائل به طور مدارم استفاده شده است.

شبکه های عصبی مصنوعی

شبکه های عصبی مصنوعی (ANN) که معمولاً به عنوان شبکه های عصبی نام برده می شوند یک الگوی ریاضی مبنی بر سیستم زیستی است. سیستم های عصبی یک الگوریتم برای بهینه سازی و یادگیری آزادانه بر اساس مفاهیم الهام گرفته از تحقیق در ماهیت مغز می باشند.

۷- روش شناسی این تحقیق

در این قسمت براساس موارد ذکر شده روش این تحقیق مشخص می شود.

روش این تحقیق بر مبنای نتیجه

پژوهش حاضر از نظر اهداف از نوع تحقیقات کاربردی است، چراکه مدل مفهومی که برا ی تجاری سازی پژوهش های علمی در این تحقیق ارائه می گردد، راهنما و دستورالعملی برای فعالیت های عملی خواهد بود.

روش این تحقیق بر مبنای هدف

پژوهش حاضر از نظر هدف، زیر مجموعه ی پژوهش های توصیفی قرار می گیرد. طبق ویژگی های پژوهش های توصیفی که پیش از این ارائه شد، پژوهش حاضر نیز به منظور کشف واقعیت های موجود یا آنچه که هست، در زمینه تشخیص جرائم الکترونیکی با استفاده از روش های داده کاوی انجام می شود.

روش این تحقیق بر مبنای نوع داده

پژوهش حاضر، پژوهش کمی می باشد، چرا که جزو تحقیقاتی می باشند که از داده های کمی جهت توصیف متغیرها متغیر وابسته: تشخیص جرائم الکترونیکی

متغیر مستقل: زمان ورود، تعداد اشتباهات، مبلغ حواله و... استفاده می کند.

۸- ماشین بردار پشتیبان (SVM)

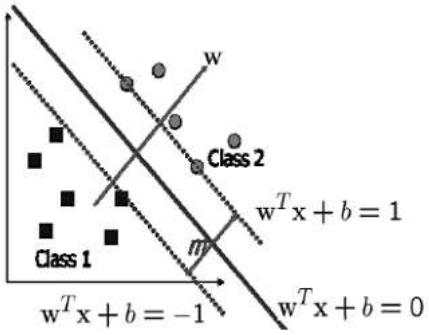
شهرت SVM به خاطر موفقیت آن در تشخیص حروف دست نویس است که با شبکه های عصبی به دقت تنظیم شده برابری می کند (۱/۱٪ خطا). هدف این دسته الگوریتم ها تشخیص و متمایز کردن الگوهای پیچیده در داده ها است (از طریق کلاسترینگ، دسته بندی، رنکینگ، پاکسازی و...). در SVM با فرض اینکه دته ها به صورت خطی جداپذیر باشند، ابر صفحه هائی با حداکثر حاشیه را به دست می آورد که دسته ها را جدا کنند. در مسایلی که داده ها به صورت خطی جداپذیر نباشند داده ها به فضای با ابعاد بیشتر نگاشت پیدا می کنند تا بتوان آنها را در این فضای جدید به صورت خطی جدا نمود. اگر دو دسته وجود داشته باشند که به صورت خطی از هم جداپذیر باشند، الگوریتم های مختلفی از جمله پرسپتون می توانند این جداسازی را انجام دهند (بریجش[1]، ۲۰۱۱). ماشین بردار پشتیبان یک طبقه بندی کننده دودویی غیر آماری است که در سال های اخیر بسیار مورد توجه قرار گرفته است تفاوت اساسی این طبقه بندی کننده با طبقه بندی کننده های آماری در این است که برای پردازش و طبقه بندی داده های ابرطیفی نیازی به کاهش تعداد باندها ندارد. در این روش با استفاده از تمامی

[1] Brijesh

باندها و یک الگوریتم بهینه سازی نمونه هایی که مرزهای کلاس ها را تشکیل می دهند به دست می آید و با استفاده از آنها یک مرز تصمیم گیری خطی بهینه برای جدا کردن کلاس ها محاسبه می شود. این نمونه ها را بردار های پشتیبان می گویند. در شکل ۱ مرز تصمیم گیری بهینه و بردار های پشتیبان نشان داده شده است. مشخص است که در بین نمونه های آموزشی همواره زیر مجموعه ای وجود دارد که بتواند بهترین مرز تصمیم گیری را تعریف کند. برای مثال نقاط آموزشی دارای کمترین فاصله از مر زهای تصمیم گیری می توانند به عنوان زیر مجموعه ای برای تعریف مرزهای تصمیم گیری در نظر گرفته شوند.

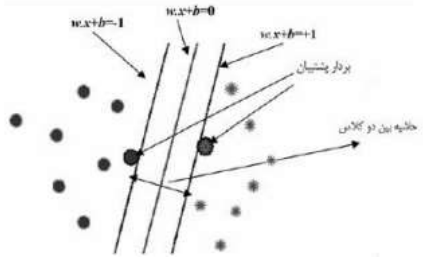
یکی از مزایای بسیار مهم ماشین بردار پشتیبان بهبود بازده این طبقه بندی کننده با افزایش بعد داده ها است. در واقع این طبقه بندی کننده یک طبقه بندی کننده آماری نیست و مستقیماً با استفاده از نقاط آموزشی مرز تصمیم گیری بین دو کلاس را به دست می آورد. توجه داشته باشید که منظور از غیر آماری بودن ماشین بردار پشتیبان این است که در این طبقه بندی کننده از ممان های آماری استفاده نمی شود و بنابراین خطای حاصل از تخمین ممانهای آماری که باعث پدیده هیوز می شود در اینجا وجود ندارد. از طرفی هر چه بعد داده ها بیشتر شود فاصله کلاس ها از یکدیگر بیشتر شده و امکان جدا سازی آنها با تعداد محدودی نقاط آموزشی افزایش می یابد. ایده SVM برای جداسازی دسته‌ها این است که دو صفحه مرزی بسازد: دو صفحه مرزی موازی با صفحه دسته بندی رسم کرده و آندو را آنقدر از هم دور می کنیم که به داده ها برخورد کنند.

صفحه دسته بندی که بیشترین فاصله را از صفحات مرزی داشته باشد، بهترین جدا کننده خواهد بود.



در صورت استفاده مناسب از SVM این الگوریتم قدرت تعمیم خوبی خواهد داشت، زیرا علیرغم داشتن ابعاد زیاد از پرهیز می کند. این خاصیت ناشی از بهینه سازی این الگوریتم است که در فشرده سازی اطلاعات استفاده می شود. در واقع به جای داده های آموزشی از بردارهای پشتیبان استفاده می کند.

در این روش با استفاده از تمامی باندها و یک الگوریتم بهینه سازی، نمونه هایی که مرزهای کلاس ها را تشکیل می دهند به دست می آورند. این نمونه ها را بردارهای پشتیبان گویند. تعدادی از نقاط آموزشی که کمترین فاصله تا مرز تصمیم گیری را دارند می توانند به عنوان زیر مجموعه ای برای تعریف مرزهای تصمیم گیری و به عنوان بردار پشتیبان در نظر گرفته می شود.



تحلیل محتوا از تکنیک های مربوط به روش اسنادی در پژوهش علوم اجتماعی میباشد. روش اسنادی به کلیه متدهایی گفته می شود که در آن ها هدف پژوهش با مطالعه، تحلیل و بررسی اسناد و متون برآورده می شود. چنانچه از نام تحلیل محتوا پیداست، این تکنیک به تحلیل و آنالیز محتوا می پردازد و می کوشد با استفاده از تحلیل داده هایی را در مورد متن استخراج کند.

۹- انواع تحلیل محتوا

در یکی از جدید ترین تقسیم های تحلیل محتوا، دوگونه کمی و کیفی ذکر می شود. روش کیفی تحلیل محتوا که گاهی استنباط نتایج بر اساس بودن یا نبودن ویژگیهایی که در پیام تعریف شده است، غالباً برای اجرای بهتر مسایل در علوم اجتماعی- کاربردی مورد توجه قرار گرفته است. حامیان فنون کیفی این فرض را مورد تردید قرار داده اند که برای نتیجه گیری فراوانی شاخص ها لزوماً نشانه اهمیت آنهاست.

۱۰- رفتارهای مشتریان

برای استفاده از قابلیت سیستم SVM در شناسایی رفتارهای مشکوک، نخست کلیهٔ رفتارهای کاربران در پنج سطح مختلف دسته بندی و سپس سیستم بردار پشتیبان بر اساس این کلاسها آموزش داده میشود. در ادامه، ابتدا شیوهٔ دسته بندی رفتارهای کاربران، تشریح و سپس به توضیح سیستم بردار پشتیبان و نحوهٔ تشخیص رفتارهای مشکوک مشتریان در سامانهٔ دیوان محاسبات کشور پرداخته خواهد شد.

همچنان که قبلاً عنوان شد پژوهشهای قبلی غالباً به شناسایی جرم بدون درنظر گرفتن شدت و ضعف آن بسنده می‌کنند. لیکن هدف اصلی در این مقاله شناسایی رفتارهای مشکوک مشتریان است. بدین لحاظ، برای رسیدن به تعریف واضحی از "کاربر مشکوک" ابتدا باید تعاریف مشخصی از رفتارهای مختلف کاربران خدمات الکترونیکی و رده های مختلف آن به وجود آید، بدیهی است این دسته بندی کمک شایانی به اتخاذ راهبرد صحیح برای نوع برخورد با هریک از آن دسته ها می نماید. از این رو پنج دسته رفتار جداگانه به شرح زیر تعریف شده است:

الف- رفتارعادی: شامل کاربرانی است که عملیات آنها به صورت عادی، بدون اشتباه و کامل انجام شده است.

ب- رفتارکمی مشکوک: شامل کاربرانی است که در هنگام ورود خطا داشته و به تلاش برای ورود غیرمجاز مظنون هستند.

ج- رفتار مشکوک: شامل کاربرانی است که در هنگام ورود خطاهای پی درپی داشته و به تلاش برای ورود غیرمجاز مظنون هستند و همچنین بدون انجام عملیات خاصی مرتباً وارد سامانه شده اند.

د- رفتار بسیارمشکوک: شامل کاربرانی است که به تلاش برای ورود غیرمجاز مشکوک و همچنین عملیات خاصی را خارج از عرف معمول تکرار کرده اند.

ه- رفتار خطرناک: شامل کاربرانی است که از مرورگرهای ناشتاخته استفاده کرده و رفتارهای یک کاربر مشکوک را نیز انجام داده اند.

در مرحلهٔ اول واسط کاربری، اطلاعات مربوط به متغیرهای ورودی سیستم را به شکل اعدادی قطعی دریافت می کند. برای تعیین پارامترهای ورودی، سیستم به طور کامل بررسی شد این بررسی از دیدگاه کاربری و سیستمی صورت گرفت. بدین منظور دو جدول شناسایی شد که اطلاعات مختلفی شامل نوع مرورگر و شمارهٔ شناسایی اتصال به اینترنت، نوع فعالیت و تراکنشی که انجام داده به همراه جزئیات آن و همچنین تاریخ و زمان انجام هر فعالیت و تراکنش از ورود تا پرداخت قبض، انجام حواله، گرفتن صورتحساب برای هر کاربر در آن ثبت شده است.

۱۱- یادگیری با نظارت

یادگیری تحت نظارت، یک روش عمومی در یادگیری ماشین است. یک مجموعه از مثال های یادگیری وجود دارد بازای هر ورودی، مقدار خروجی و یا تابع مربوطه نیز مشخص است. هدف سیستم یادگیر بدست آوردن فرضیه ای است که تابع و یا رابطه بین ورودی و یا خروجی را حدس بزند به این روش یادگیری با نظارت گفته میشود. به این معنی که یک شخص ناظری وجود دارد که برچسب گذاری برای تمایز دسته های مختلف را بر روی اسناد اعمال می کند.

۱۲- دسته بندی اسناد

دسته بندی اسناد یا رده بندی اسناد یک مسئله مهم در علوم مرتبط با کتابخانه ها و همچنین علوم کامپیوتر است. کار اصلی دسته بندی این است که اسناد را به یک یا چند طبقه یا رده تقسیم کند. این کار ممکن است به صورت دستی یا به صورت الگوریتمی انجام شود. رده بندی به صورت دستی مانند کاری است که در یک کتابخانه انجام می شود و کتابها به دسته های مختلف تقسیم می شوند. روش الگوریتمی رده بندی اصولا در علم کامپیوتر و در مبحث داده کاوی جای می گیرد و به صورت یادگیری ماشین انجام می شود. اسناد ممکن است بر اساس موضوع و یا براساس یک ویژگی خاص (مانند نام نویسنده) رده بندی شوند. در رده بندی موضوعی اسناد ابتدا اسنادی با موضوع معین به عنوان نمونه برای یادگیری به ماشین داده می شود و ماشین با توجه به کلمات داخل هر سند به یادگیری می پردازد و با در نظر گرفتن احتمال وجود کلمات به پیش بینی موضوعی سندهای دیگر می پردازد. هدف این دسته الگوریتم ها تشخیص و متمایز کردن الگوهای پیچیده در داده هاست از طریق کلاسترینگ، دسته بندی، رنکینگ، پاکسازی و غیره (کالایکوماران[1]، ۲۰۱۲). مبنای کاری دسته بندی کنندهٔ SVM دسته بندی خطی داده ها است و در تقسیم خطی داده ها سعی میکنیم خطی را انتخاب کنیم که حاشیه اطمینان بیشتری داشته باشد. با فرض اینکه دسته ها بصورت خطی جداپذیر باشند، ابرصفحه هائی با حداکثر حاشیه را بدست می آورد که دسته ها را جدا کنند. اگر دو دسته وجود داشته باشند که بصورت خطی از هم جداپذیر باشند، بهترین جدا کننده این دو دسته چیست. ایده SVM برای جداسازی دسته ها به این صورت است که ابتدا یک صفحه دستهبندی بسازید. دو صفحه مرزی موازی با صفحه دسته بندی رسم کرده و آندو را آنقدر از هم دور میکنیم که به داده ها برخورد کنند. صفحه دسته بندی که بیشترین فاصله را از صفحات مرزی داشته باشد، بهترین جدا کننده خواهد بود. نزدیکترین داده های آموزشی به ابر صفحه های جدا کننده بردار پشتیبان نامیده میشوند. فشرده سازی اطلاعات در الگوریتم SVM به این صورت است که بجای داده های آموزشی از بردارهای پشتیبان استفاده می کند. در مسایلی که داده ها بصورت خطی جداپذیر نباشند داده ها به فضای با ابعاد بیشتر نگاشت پیدا میکنند تا بتوان آنها را در این فضای جدید بصورت خطی جدا نمود.

۱۳- نتیجه گیری

بی تردید شرایط اجتماعی بشر، رویارویی با پدیده ای به نام جرم را غیر قابل اجتناب می سازد و انسان همواره نیازمند دانش تحلیل جرم است. تحلیل جرم عبارت است از به کارگیری شیوه ای نظام مند جهت شناسایی، کشف و پیش بینی جرایم ورودی یک سیستم تحلیل جرم، داده ها و اطلاعات منتسب به متغیرهای جرم است و خروجی آن پاسخ به پرسش های تحلیلی، استخراج دانش و در نهایت مصور سازی نتایج است. جرایم ناهنجاری های اجتماعی هستند که جوامع به گونه های مختلف هزینه های زیادی برای آن می پردازند شناسایی الگوهای جرم و کشف جرایم از دیرباز مورد توجه سازمان های قضایی و پلیسی بوده است. در گذشته از ابزارهای مختلفی برای تحلیل جرایم و کشف جرم و شناسایی روابط پیچیده میان جرم و مجرم استفاده می کردند. کشف جرائم رایانه ای به معنای لغوی استفاده از یک سری روش های هدفمند از تکنیک ها و دستورالعمل ها برای بدست آوردن و جمع آوری آثار جرم از تجهیزات و سیستم های اطلاعاتی و محاسباتی و انواع دستگاه های ذخیره سازی اطلاعات و رسانه های دیجیتال است، به گونه ای که بتوان آنها را در یک قالب منسجم و قابل ارائه و مفهوم در دادگاه جرائم رایانه ای ارائه کرد؛ اما در واقع به فرآیند نگهداری، شناسایی، استخراج، تفسیر و مستند سازی آثار و شواهد کامپیوتری که دارای ویژگی های قانونی یک اثر جرم از قبیل انجام شدن فرآیند های قانونی، صحت و تمامیت اثر جرم، گزارش دهی درست اطلاعات بدست آمده و در نهایت دریافت رای نهایی از یک دادگاه قانونی یا یک سیستم قانونی معتبر می باشند کشف جرائم رایانه ای گفته می شود. در حقیقت کشف جرائم رایانه ای یک علم است که به آن عمل دریافت، پردازش و شناسایی داده ها از کامپیوتر ها به روشی است که تمامی آثار بدست آمده در دادگاه جرائم رایانه ای قابل ارائه باشند.

[1] Kalaikumaran

امروزه داده کاوی می تواند به عنوان یک ابزار نوین به کار گرفته شود تا مشکلات و موضوعات شناسایی جرم را مدل نماید. در دهه اخیر داده کاوی به عنوان ابزاری قدرتمند برای سازمان های پلیسی و جنایی مطرح شده و به کار گرفته می شود. در ایران نیز با توجه به راه اندازی سیستم های مکانیزه و ثبت اطلاعات مجرمان و وقایع رخ داده در سسیستم پلیس ایران، به نظر می رسد داده کاوی و ابزارهای آن می تواند نقش مؤثری در شناسایی سریع الگوهای جرم و کشف جرایم داشته باشد.

داده کاوی عبارت است از اقتباس یا استخراج دانش از مجموعه ای از داده ها، به دیگر بیان، داده کاوی فرایندی است که با استفاده از روشهای هوشمند، دانش را از مجموعه ای از داده ها استخراج می کند. گستره وسیع کاربردهای داده کاوی این موضوع را به عنوان یکی از مهمترین حوزه های پژوهشی مطرح نموده است. یکی از عرصه های مهم کاربرد داده کاوی، حوزه جرم شناسی است، جرم شناسی به فرایند تحلیل و آنالیز یک جرم و شناخت ویژگی های آن اطلاق می شود. در واقع جرم شناسی شناخت جزئیات یک جرم و روابط نامحسوس آن با مجرم را شامل می شود. حجم بسیار زیاد داده های مربوط به جرم ها و مجرمین از یک سو و وجود روابط معنایی پیچیده و نامحسوس میان این اطلاعات از دیگرسو، جرم شناسی را به یکی از مهمترین حوزه های کاربردی داده کاوی مبدل نموده است. با استخراج ویژگی های جرم، گام ابتدایی برای هر گونه تحلیل و بررسی کارشناسانه روی ویژگیهای یک جرم امکان پذیر خواهد بود. در واقع دانش حاصل از اعمال روش های داده کاوی در حوزه جرم شناسی بستر مناسبی را برای پشتیبانی اطلاعاتی کارشناسان و انجام فعالیت های آتی پلیس فراهم می آورد. تا کنون تلاش زیادی توسط محققان در رشته های مختلف برای کشف دانش مخفی در مجموعه داده ها صورت گرفته است. در اواخر سال های ۱۹۸۰ اصطلاح جدید، کشف دانش در پایگاه داده به کار برده شد و به سرعت از سوی متخصصان هوش مصنوعی و فراگیری ماشین برای پوشش فرایند کلی استخراج دانش از پایگاه های داده از بدو تعریف مسئله و اهداف آن تا تحلیل های نهایی نتایج مورد استفاده قرار گرفت.

در این زمینه واژه داده کاوی به عنوان یک مرحله یا گام در فرایند کشف دانش مورد استفاده قرار گرفت. این تفسیر توسط فیلد در سال ۱۹۹۴ بدین شکل ارائه شد یعنی فرایند کشف دانش که شامل مجموعه ای از شیوه های ریاضی و رایانه است که براساس تحلیل داده های موجود در یک پایگاه داده بزرگ، برای یافتن راه حلی براساس الگوهای کشف شده در داده ها و به کاربردن آن برای مسئله تعریف شده است به عبارت دیگر کشف دانش در پایگاه داده فرایند شناسایی درست، ساده، مفید و همچنین کشف الگوها و مدل های قابل فهم در داده هاست. داده کاوی مرحله ای از فرایند کشف، دانش است و شامل الگوریتم های مخصوص به خود است، به طوری که الگوها و یا مدل ها را در محدودیت های مؤثر محاسباتی قابل قبول کشف می کند. منظور از الگوی مفید، مدلی معتبر، ساده، قابل فهم و جدید در داده هاست که ارتباط میان زیر مجموعه ای از داده ها را توصیف می کند. در واقع داده کاوی یکی از مهم ترین روش هایی است که به وسیله آن الگوهای مفید در درون داده ها با حداقل دخالت کاربران شناخته می شود و اطلاعاتی را در اختیار کاربران و تحلیلگران قرار می دهد تا بر اساس آن ها تصمیمات مهم و حیاتی در سازمان ها اتخاذ شود (کورپوریشن[1]، ۱۹۹۴).

در ویرجینای غربی حدود ۹۰۰ مرکز اجرای قانون وجود دارد که WVSPFL یکی از این مراکز در جنوب ویرجینای غربی است. این مرکز امکاناتی را جهت تحلیل نمونه های مربوط به حوادث ویژه فراهم کرده است. در واقع آزمایشگاه جنایی ایالت ویرجینای غربی از یک ابزار نرم افزاری به نام FIMS (سیستم مدیریت اطلاعاتی جنایی) استفاده کرده است. این ابزار با به کارگیری اصول آمار و الگوریتم های داده کاوی می تواند داده های جرم را تحلیل نماید و نتایج منطقی را به دست آورد. نرم افزار FIMS از سه قسمت تشکیل شده است که عبارتنداز:

آمار حوادث، داده کاوی و تصویرسازی داده. با کمک این ابزار، آمار حوادث به مراکز اجرای قانون ارائه می شود و آنان را در تصمیم گیری مربوط به بررسی و کنترل حوادث، راهنمایی می کند برای ساخت یک مدل تحلیلی درست باید داده ها را از جداول مختلف جمع آوری کرد. این ابزار به کاربر امکان می دهد تا اطلاعات مرتبط را از پایگاه داده های مختلف به وسیله دستورات SQL جمع آوری کند.

[1] Corporation

همچنین این ابزار می تواند در شناسایی الگوهای جرم در بین داده حوادث ذخیره شده در پایگاه داده FIMS (جرم، مظنون و اطلاعات قربانی) به آزمایشگاه جنایی پلیس در ایالت ویرجینای غربی کمک کند. این ابزار به WVSPFL و دیگر مراکز اجرای قانون در بهره برداری بهتر منابع و پیش بینی فعالیت جنایی ممکن کمک خواهد کرد.

چگونه تکنیک داده کاوی می تواند به کارایی مأموران تحقیق و بررسی صحنه جرم کمک کند. هوش قانونی (مانند اثر انگشت ها با شناسایی DNA) به عنوان یک روش قانونی استاندارد برای تحقیق و کشف یک طیف وسیع از انواع جرم اهمیت دارد. پلیس Nor thamptonshire تمام جرایم را درون یک پایگاه داده رابطه ای Oracle ثبت کرد. بخش پشتیبان علمی از سیستم رایانه Trak- X در ثبت و مدیریت همه توابع استفاده کرد (Trak- X به صورت داخلی طراحی شده بود و توسط یک فروشنده نرم افزار توسعه یافته بود).

بیشتر کارشناسان معتقدند هنگامی که سرقت های مسلحانه همراه با تهاجم های خطرناک است، خسارت وارد شده به قربانی وخیم تر است. بنابراین داده سرقت مسلحانه را در ریچموند، ویرجینا، در یک دزدی احتمالی آزمایش کردند تا خشونت های مسلحانه را محدود کنند. با استفاده از Clementine SPSS یک مدل توسعه دادند که عوامل وابسته به یک احتمال فزاینده از یک سرقت مربوط به حملات مسلحانه را مشخص کرد سپس این نتایج را در نقشه ای برای استفاده توسط نیروهای گشتزنی و واحد تاکتیکی گسترش دادند. نتایج نشان داد که مناطق با ریسک های بالا مشابه هم هستند، اما دزدی های مسلحانه به صورت یکسان در همه جا توزیع نشده است. در ناحیه های مشخص شده، سرقت های مسلحانه لزوماً به یک درجه افزایش یافته وابسته نبود، بلکه آن ها به یک احتمال فزاینده وابسته بودند. این یافته ها اختلاف های دقیق بین بسامد نسبی از جرم و احتمال تهدیدهای جدی را مشخص کرد. مناطق با ریسک بالا شامل، مناطق جغرافیایی کوچکتری در شهر می شدند. با تعیین مناطق خطر خیز امکان افزایش گشت زنی و گسترش واحدهای تاکتیکی در زمان بسیار کوتاه به صورت هدفمند فراهم شد. به طور خلاصه می توان نتیجه گرفت:

جرایم ناهنجاری های اجتماعی هستند که جوامع هزینه زیادی را به گونه های مختلف بابت آن می پردازند. مهم ترین هدف های دولت ها به کارگیری یک سیاست جنایی کارآمد، از بین بردن فرصت های ارتکاب جرم، پیشگیری وضعی از وقوع جرم و برخورد با مجرمان است. در این راستا سازمان های پلیسی از شیوه های مختلفی مانند (افزایش نیروهای پلیس، دوربین های مدار بسته و. ...) استفاده کردند. با توسعه و گسترش فناوری اطلاعات در سازمان ها و ایجاد وب سایت پورتال دیوان محاسبات کشور های اطلاعاتی، داده کاوی نیز به عنوان یک ابزار نرم افزاری قدرتمند و به مراتب کم هزینه تر و کارآمدتر در اختیار سازمان های پلیس قرار گرفته است. با استفاده از داده کاوی می توان الگوهای جرم را شناسایی کرد تا بتوان جرایم را پیش بینی کرد و از وقوع آن پیشگیری نمود.

در دنیا ابزارهای مختلفی برای تحلیل داده های جرم به کار گرفته شده است که برخی از آن ها عبارتند از:

Brain Maker, CrimeStat III CART, Crime Connect. Crime Point Web. BRAINCEL. AC

برای استفاده از این ابزارها باید فرایندهای اصلی داده کاوی را در چارچوبی معین طی کرد تا بتوان به مدل های دقیق تری برای جرایم و کشف الگوهای ناشناخته در وب سایت پورتال دیوان محاسبات کشور های اطلاعاتی دست یافت. بدین ترتیب پلیس می تواند به صورت مؤثر و کارآمد نیروهای خود را در مناطق به کار گرفته و با اقتدار به اداره و کنترل حوزه استحفاظی خود بپردازد.

همان طور که در قسمت های قبل اشاره شد داده کاوی دارای قابلیت ها و امکاناتی است که برخی از آنها عبارتند از:

۱. کشف ارتباط بین داده ها: از مهم ترین قابلیت های داده کاوی شناسایی و کشف ارتباطات ناشناخته در وب سایت پورتال دیوان محاسبات کشور اطلاعاتی است که به دست آوردن آن برای کاربران تقریباً غیر ممکن است.

۲. امکان تصویر سازی داده ها به کمک GIS و روش های داده کاوی می توان نقاط حادثه خیز و خطرناک و همچنین توزیع جغرافیایی جرم را بر حسب سن، جنسیت، مکان و نوع جرم به صورت نمودار نمایش داد و راهبرد سازمان را تدوین کرده و یا در صورت لزوم تغییر داد.

۳. تحلیل حجم بسیار زیاد داده ها: به کمک روش های داده کاوی می توان داده های فراوان موجود در وب سایت پورتال دیوان محاسبات کشور های اطلاعاتی را تحلیل کرد که این کار به صورت دستی و بدون کمک گرفتن از روش‌های داده کاوی سخت و دشوار و زمانبر امکان پذیر است.

در کنار این قابلیت ها نقدهایی به داده کاوی جرم وارد است که برخی از آنها عبارتند از:

۱. در صورت وجود داده های شلوغ و کثیف (پالایش نشده) در وب سایت پورتال دیوان محاسبات کشور اطلاعاتی، عملا بهره برداری از ابزارهای داده کاوی بیهوده است.

۲. به منظور استفاده از الگوریتم های داده کاوی باید افراد آموزش های تخصصی موردنیاز را در این حوزه دیده باشد.

۳. استفاده از روش های داده کاوی در دنیای واقعی منوط به انجام آزمایش های پیچیده در محیط آزمایشگاهی و مقایسه آن با نتایج استخراج شده به صورت دستی توسط کارشناسان خبره جهت تایید صحت و دقت پیش بینی روش هاست. (کیوان پور۱۳۸۷)

امروزه تشخیص جرم و بهبود سطح امنیت در جرائم الکترونیکی بسیار مهتر از گذشته است؛ بنابراین در این پژوهش ما رفتار مشتریان وب سایت پورتال دیوان محاسبات کشور را بررسی کردیم که به پنج دسته رفتار جداگانه به شرح زیر تعریف شده است:

الف - رفتارعادی: شامل کاربرانی است که عملیات آنها به صورت عادی، بدون اشتباه و کامل انجام شده است.

ب - رفتارکمی مشکوک: شامل کاربرانی است که در هنگام ورود خطا داشته و به تلاش برای ورود غیرمجاز مظنون هستند.

ج - رفتار مشکوک: شامل کاربرانی است که در هنگام ورود خطاهای پی درپی داشته و به تلاش برای ورود غیرمجاز مظنون هستند و همچنین بدون انجام عملیات خاصی مرتباً وارد سامانه شده اند.

د - رفتار بسیارمشکوک: شامل کاربرانی است که به تلاش برای ورود غیرمجاز مشکوک و همچنین عملیات خاصی را خارج از عرف معمول تکرار کرده اند.

ه - رفتار خطرناک: شامل کاربرانی است که از مرورگرهای ناشتاخته استفاده کرده و رفتارهای یک کاربر مشکوک را نیز انجام داده اند.

سپس انواع رفتارهای متفاوت براساس داده کاوی اطلاعات مشتریان چند کلاس برای این کاربران وب سایت پورتال دیوان محاسبات کشور تعریف کردیم که و با استفاده از ماشین بردار پشتیبان (SVM) آنها را کلاس بندی نمودیم (بریل[1]، ۲۰۰۸).

منابع

۱. جهرمی، وحید، نوری مطلق، محمد، هوشیار، عبد اله، (۱۳۹۰)، جرایم الکترونیکی و چالش های اجرای قوانین مربوط به آن در جامعه، همایش منطقه ای چالش های جرایم رایانه ای در عصر امروز.

۲. روحانی رانکوهی، سید محمد تقی، (۱۳۸۰). مفاهیم بنیادی پایگاه داده ها. چاپ اول. انتشارات جلوه.

۳. ژیاوی هان، میشلین کمبر، جیان پی (۱۳۹۲)، داده کاوی؛ تکنیک ها، نسترن حاجی حیدری، سیدبهنام خاکباز، انتشارات دانشگاه تهران.

۴. کاظمی، پروانه، حسین پور، جواد، (۱۳۸۸)، کاربرد داده کاوی در سازمان های پلیسی و قضایی به منظور شناسایی الگوهای جرم و کشف جرایم، نشریه کارآگاه، دوره دوم، شال دوم، شماره ۸.

۵. کیوان پور، محمدرضا، پهلوان زاده، افسانه، سجودی، مهدیه، جاویده، مصطفی، (۱۳۸۷)، تحلیل و کشف جرم با بهره گیری از روش های داده کاوی، دومین کنفرانس داده کاوی ایران.

6. Bickel, S, & Scheffer, T. (2007). Dirichlet-enhanced spam filtering based on biased samples. Advances in neural information processing systems, 19, 161.2010

[1] Bryl

7. Bose, I, & Mahapatra, R. K. (2001). Business data mininga machine learning perspective. Information & management, 39(3), 211-225.
8. Brijesh Kumar Baradwaj Research Scholor, Singhaniya University, Rajasthan, India Saurabh Pal Sr. Lecturer, Dept. of MCA,VBS Purvanchal University, Jaunpur-222001, India "Mining Educational Data to Analyze Students, Performance ".2011
9. Brown, D. E. (1998, October). The regional crime analysis program (RECAP): a framework for mining data to catch criminals. In Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on (Vol. 3, pp. 2848-2853). IEEE.
10. Bryl, A. (2008). DISI-University of Trento Using Locality for Improving SVM based Spam Filtering.
11. Chen, H, Chung, W, Xu, J. J, Wang, G, Qin, Y, & Chau, M. (2004). Crime data mining: a general framework and some examples. Computer, 37(4), 50-56.
12. Chitra, S, Jayanthan, K. S, Preetha, S, & Shankar, R. U. (2012). Predicate based Algorithm for Malicious Web Page Detection using Genetic Fuzzy Systems and Support Vector Machine. International Journal of Computer Applications, 40(10), 13-19.
13. Corporation. T.C (1999) "Introduction to Data Mining and Knowledge Discovery ", Third Edition ISBN: 1-892095-02-5.
14. Dharwa, J. N, & Patel, A. R. (2011). A Data Mining with Hybrid Approach Based Transaction Risk Score Generation Model (TRSGM) for Fraud Detection of Online Financial Transaction. International Journal of Computer Applications,16(1), 18-25.
15. Fdez-Riverola, F, Iglesias, E. L, Díaz, F, Méndez, J. R, & Corchado, J. M. (2007). Applying lazy learning algorithms to tackle concept drift in spam filtering. Expert Systems with Applications, 33(1), 36-48.
16. Frawley, W. J, Piatetsky-Shapiro, G, & Matheus, C. J. (1992). Knowledge discovery in databases: An overview. AI magazine, 13(3), 57.
17. Hajian, S, Domingo-Ferrer, J, & Martínez-Ballesté, A. (2011, April). Discrimination prevention in data mining for intrusion and crime detection. InComputational Intelligence in Cyber Security (CICS), 2011 IEEE Symposium on(pp. 47-54). IEEE.
18. Han, Q, Zeng, L, Liu, Y, Liu, Y, An adaptive clustering algorithm for road abnormal region analysis, Transactions of the Institute of Measurement and Control July 3, 2013.
19. Johnson, T. A. (Ed.). (2013). Forensic computer crime investigation. CRC Press.
20. Kalaikumaran, T, & Karthik, S. (2012). Criminals and crime hotspot detection using data mining algorithms: clustering and classification. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET),1(10), pp-225.

مطالعات آینده پژوهی و سیاستگذاری

دوره ۳، شماره ۲، تابستان ۱۳۹۶، صفحات ۷۷-۶۵

Investigation and Identification of the Proposed Algorithm for Electronic Crimes in the Website of the Supreme Audit Court of Iran

Mina Jantani

B.Sc. in computer software, Audit Court of Kerman Province, Kerman, Iran

Abstract

The electronic and virtual offenses have a short history and have appeared only since the last 20 years. Actually, the simplified use of computers for all people and the reduced price of access to information technology tools have given rise to a new era of "electronic crimes" in virtual space and have challenged the security, regulatory, and oversight institutions. Electronic crimes which are related to the third-generation crimes of the computer and the Internet occur in virtual environments or electronic spaces. In the meantime, the scientific and intelligent methods of crime detection have attracted the attention of many criminologists due to their scientific and mathematical support. One of these strategies is data mining, which refers to the process of extracting knowledge from a set of data using intelligent methods. In this paper, using content and document analysis and the support vector machine, we designed an intelligent system to detect electronic crimes in a virtual environment. We tested the proposed algorithm to identify the behavior of the users of the Supreme Audit Court of Iran's website and to diagnose suspicious behaviors with varying degrees of severity. To this end, we categorized all users' behaviors using data mining and content analysis method at five different levels. We then tested the vector support system based on these classes.

Keywords: data mining, electronic crimes, crime, the Supreme Audit Court of Iran