ماهنامه علمی تخصصی پایا شهر

شماره مجوز مجله:۸۰۱۴۰۰ زمان پذیرش نهایی:۱۴۰۰/۰۲/۱۵

تکنیک های تشخیص جرم با استفاده از داده کاوی و یادگیری ماشین

احسان نریمانی[۱*]، افشین رضاخانی[۲]

[۱] کارشناس ارشد نرم افزار، دانشگاه غیرانتفاعی یاسین بروجرد، بروجرد، ایران
[۲] استادیار و عضو هیئت علمی دانشگاه آیت اله بروجردی، بروجرد، ایران

چکیده

یکی از پیچیده ترین مسائل در شهرهای جهان، نرخ بالای جرم و افزایش ناهنجاریهای اجتماعی در آنهاست. بروز انواع جرم و ناهنجاری باعث ایجاد حس ناامنی و تحمیل مشکلات مالی بر دوش جامعه، دولت و تشکیلات قضایی کشور میشود. ویژگی‌های جرم و جنایت در یک پایگاه داده به مانند یک معدن طلا ارزشمند هستند. کشف دانش پنهان در این پایگاه‌های داده کلید حل مسئله است، روش‌های و تکنیک های داده کاوی به عنوان ابزارهای کاوش مخازن داده به استخراج دانش نهفته در دل آنها می پردازد. در حال حاضر مطالعات خوبی در امر داده کاوی اطلاعات اعمال مجرمانه در دنیا انجام شده است اما متاسفانه در داخل کشور فعالیت چندانی به چشم نمی خورد. در دنیای مدرن امروز به موازات گسترش شیوه های نوین زندگی نوع و شکل ارتکاب جرایم نیز تغییریافته است و مجرمین شگردها و روش های جدیدی را برای ارتکاب و پنهان نمودن آثار جرم ارتکابی و به دنبال آن هویت خود انتخاب میکنند. در این پژوهش قصد داریم با ارائه پیشنهاداتی برای سیستم گزارشات جرائم، بر میزان تسریع کشف جرم تاثیر گذار بگذاریم.

کلیدواژه: تشخیص جرم، داده کاوی، یادگیری ماشین.

مقدمه

کشف دانش در پایگاه داده فرایند شناسایی درست، ساده، مفید و نهایتا الگوها و مدل‌های قابل فهم در داده ها می‌باشد. داده‌کاوی، مرحله‌ای از فرایند کشف دانش می‌باشد و شامل الگوریتم‌های مخصوص داده‌کاوی است، بطوری که تحت محدودیت‌های مؤثر محاسباتی قابل قبول، الگوها یا مدل‌ها را در داده کشف می کند. به بیان ساده‌تر، داده کاوی به فرایند استخراج دانش ناشناخته، درست و بالقوه مفید از داده اطلاق می شود. تعریف دیگر این است که، داده‌کاوی گونه‌ای از تکنیک‌ها برای شناسایی اطلاعات و یا دانش تصمیم گیری از قطعات داده می‌باشد، به نحوی که با استخراج آن ها، درحوزه‌های تصمیم‌گیری، پیش‌بینی، پیش گویی، و تخمین مورد استفاده قرار گیرند. به این دلیل اغلب به داده کاوی، تحلیل داده ای ثانویه[1] گفته می‌شود[۳]. بطور کلی داده کاوی به معنای، تحلیل داده و کشف الگوهای پنهان با استفاده از ابزارهای خودکار و یا نیمه خودکار است. پیشگیری از وقوع جرم یکی از اهداف کلیدی در حوزه مدیریت کلان هر کشوری می باشد. در حوزه های نظامی و انتظامی هزینه های پیشگیری از وقوع جرم بسیار کمتر از هزینه مقابله با وقوع جرم است در صورتی که برای پیشگیری از وقوع جرایم باید اطلاعات لازم در زمینه جرایم و تجزیه و تحلیل اطلاعات در دسترس باشد. در حوزه های انتظامی و امنیتی همواره ما با حجم بسیار بالایی از اطلاعات سرو کار داریم که تجزیه و تحلیل چنین حجمی از اطلاعات با روشها و متدهای سنتی بسیار مشکل است که برای اینکار از روش‌های نوین داده کاوی استفاده می کنیم[۴]. مبارزه با جرم و کجرویهای اجتماعی منوط به شناخت عوامل جرمزا است که با از بین بردن این عوامل یا کاهش اثرات آن میتوان از بروز جرایم پیشگیری نموده و یا از میزان آن در جامعه کاست. آمارهای موجود در کشورهای جهان نشان میدهد پیشگیری از جرم از طریق افزایش نیروهای پلیس، تدابیر شدید امنیتی، صدور احکام و مجازات های شدیدتر و احداث زندان‌های بیشتر راه به جایی نمی برد[۳]. ویژگی های بزهکار، نحوه انجام عمل مجرمانه، ویژگی های بزه دیده و رابطه بین بزهکار و بزه دیده همگی جزو پارامترهای بسیار متنوع و گوناگون دخیل در بحث تحلیل جرم و جنایت هستند که استفاده از تکنیک‌های مطرح در علوم مختلف را می طلبد[۲]. پلیس پیشگیرانه، یکی از مفاهیم در حال رشد در گرایش استراتژیکی نیروی پلیس در تمام دنیاست؛ که می تواند با صرف زمان کوتاهتری، به پیش بینی جرم، تشخیص جرم و جلو گیری از آن بپردازد و از طرف دیگر توسعه سیستم های کامپیوتری و پیشرفت تکنولوژی هوش مصنوعی در زمینه داده کاوی امکان تحلیل این حجم انبوه داده ها را با استفاده از ماشین در اختیار پلیس قرار می دهد. کشف دانش پنهان در این پایگاه‌های داده کلید حل مسئله است، روش‌های و تکنیک های داده کاوی به عنوان ابزارهای کاوش مخازن داده به استخراج دانش نهفته در دل آنها می پردازد. در حال حاضر مطالعات خوبی در امر داده کاوی اطلاعات اعمال مجرمانه در دنیا انجام شده است اما متاسفانه در داخل کشور فعالیت چندانی به چشم نمی‌خورد [۷]. در بسیاری از رسیدگی های جنایی و پرونده های مرموز و پیچیده، بررسی دقیق در تحقیقات مقدماتی به ویژه صحنه جرم منجر به یافتن مدارک و شواهدی شده است که ضمن روشن نمودن حقیقت جنایت هویت مرتکب را نیز آشکار کرده است. با توجه به این تعریف و همچنین چالش های مطرح شده ضرورت استفاده از تکنیک های داده کاوی در تحلیل علمی جرایم نمایان می گردد. فرضیاتی که در این پژوهش مورد نقد و بررسی قرار میگیرد به شرح زیر می باشد:

- استفاده از روش داده کاوی و یادگیری ماشین در تشخیص و تجزیه و تحلیل تکنیک های تشخیص جرم موثر است.
- استفاده از روش داده کاوی و یادگیری ماشین در افزایش دقت و کارایی بیشتر طبقه بندی و تشخیص جرم موثر است.

[1] Secondary Data Analysis

- روش پیشنهادی می تواند جهت تشخیص و پیش بینی جرم در اختیار نیروی پلیس قرار بگیرد.

داده‌کاوی تلاش برای استخراج دانش از انبوه داده‌های موجود است. داده‌کاوی به کمک مجموعه‌ای از روش‌های آماری و مدل-سازی، می‌تواند الگوها و روابط پنهان موجود در پایگاه‌های داده را تشخیص دهد. تاکنون ابزارها و روش‌های مختلف برای پردازش اطلاعات ساخت‌یافته توسعه داده شده است که در نتیجه آنها ساخت پایگاه‌های اطلاعاتی و ایجاد انبارهای داده به سادگی صورت می‌گیرد. امروزه سازمان‌ها قادرند با هزینه کم اطلاعات وسیعی از وضعیت کسب و کار خود جمع و نگهداری کنند و این موجب شده است که استفاده از روش‌های داده‌کاوی، ارزش قابل توجهی را برای سازمان، به دست آورد. رویکردهای موجود به مساله داده‌کاوی متنوع است. الگوریتم‌های طبقه بندی برای داده‌های حاصل از تشخیص جرم تحلیل شده و نتایج حاصل از پیاده سازی الگوریتم های طبقه بندی مورد ارزیابی قرار می گیرد. در این مطالعه سعی شده علاوه بر مقایسه قدرت پیش بینی درخت تصمیم گیری، مدلهای درخت داده کاوی با الگوریتمهای آموزش مختلف نیز مقایسه شود. انتخاب الگوریتم آموزش مناسب یکی از مهمترین مراحل طراحی یک درخت تصمیم است. با در نظر گرفتن مواردی مانند سرعت همگرایی، میزان تحت تاثیر قرار گرفتن توسط داده های نویز و حساسیت به خطای آموزش، انواع مختلفی از الگوریتم پس انتشار خطا مورد مقایسه قرار گرفتند [۱۲]. از آن جائیکه این پژوهش قصد دارد با ارائه پیشنهاداتی برای سیستم گزارشات جرائم ، بر میزان تسریع کشف جرم تاثیر گذار باشد و از طرفی به استفاده از راه حل سیستمی برای حل مسئله توجه دارد، از نوع تحقیق کاربردی می باشد. چنانچه این مجموعه تحقیقاتی از لحاظ روش تحقیق بررسی گردد ، با توجه به این امر که این تحقیق آنچه را که هست توصیف و تفسیر می کند، به شرایط و روابط موجود مربوط به دسته بندی جرائم بر اساس متغیرها می پردازد، و مجموعه ای منسجم و منظمی از داده ها را جمع آوری می کند، از نوع تحقیق توصیفی ، موردکاوی می باشد . چنانچه این تحقیق از نظر نوع داده ایی بررسی گردد، به دلیل بیان داده ها بشکل اعداد و ارقام ، از نوع کمی خواهد بود. این پژوهش در سال های ۱۳۹۸ و ۱۳۹۹ انجام گردیده است که برای انجام این پژوهش از مقالات مختلفی استفاده شده است که به تعدادی از آنها و پایگاه های دسترسی اشاره می کنیم بصورت مختصر به شرح کار آنها می پردازیم.

کارهای انجام شده

از پژوهش لنگری زاده و اروجی در سال (۱۳۹۶) با عنوان شناسه زدایی پرونده الکترونیک سلامت با استفاده از الگوریتم های یادگیری ماشین،که یک پژوهش مروری نظام مند بوده و در سال های ۲۰۱۶-۲۰۰۶ در پایگاه‌های PubMed و Science directانجام شده است مورد استفاده قرار گرفت [۵]. پژوهش عباسی راد و مدرکی که در سال (۱۳۹۶)، با عنوان ارائه راهکاری جهت تشخیص جرائم در داده های بزرگ با استفاده از داده کاوی از نوع پژوهش های خوشه بندی شده مبتنی بر چگالی برای تجزیه و تحلیل داده های جرم و جنابت مورد استفاده قرار گرفت. این پژوهش در سامانه ایرانداک نمایه شده و قابل استفاده می باشد [۲]. در پژوهش مانیان و همکاران در سال (۱۳۹۵)، که به بررسی طراحی الگوی داده کاوی پیشنهادی به منظور شناسایی مجرمان پرداختند. در این پژوهش که با بهره گیری از الگوریتم های داده کاوی به تحلیل داده های ثبت شده در بانک اطلاعاتی پلیس مربوط به دستگیرشدگان توسط گشت های انتظامی تهران بزرگ در سه ماهه اول سال ۱۳۸۹ پرداخته شده و با استفاده از آنها، الگویی طراحی شده که به شناسایی مجرمان واقعی از بین انبوه متهمان دستگیر شده اقدام کند. این پژوهش در سامانه نور مگز در دسترس بوده و با ۸ مقاله دیگر از لحاظ داده کاوی همخوانی دارد [۶]. در پژوهش ابراهیمی و همکاران در سال (۱۳۹۴)،

که به بررسی جامعیت مجموعه جرائم به منظور پیش بینی و شناسایی جرائم با استفاده از تکنیک های داده کاوی پرداخته اند که این مقاله در پایگاه مقالات ISC مورد دسترسی و مطالعه قرار گرفت [۱]. در پژوهش یه[۲] و همکاران در سال (۲۰۱۸)، که با عنوان شناسایی کلاهبرداری کارت اعتباری با استفاده از یادگیری ماشینی به عنوان تکنیک داده کاوی در پایگاه مقالات IEEE مورد مطالعه قرار گرفت و در بازه زمانی ۲۰۱۸ تا ۲۰۱۹ نمایه شد[۱۳]. در مقاله باسیونی[۳] و همکاران در سال (۲۰۱۸)، به بررسی دسته بندی ایمیل های هرزنامه و HAM با استفاده از تکنیک های یادگیری ماشین پرداخته شد. این پژوهش در پایگاه مقالات IET قابل دسترسی بوده است[۸]. در پژوهش چاکرا[۴] و همکاران در سال (۲۰۱۵)، به بررسی تکنیک های داده کاوی برای تجزیه و تحلیل داده های ساخت یافته و غیر ساخت یافته‌ی مربوط به قتل پرداختند که این پژوهش در بازه زمانی سال ۲۰۱۵ نمایه گذاری شده است[۹]. در پژوهش کلندن[۵] و همکارش در سال (۲۰۱۵)، با استفاده از الگوریتم های یادگیری ماشینی به تجزیه و تحلیل داده های جرم و جنایت پرداختند. این پژوهش درپایگاه مقالات Elsevier نمایه شده است و مورد استفاده قرار گرفت[۱۱]. در پژوهش گنزالز[۶] و همکارش در سال (۲۰۱۳)، که به بررسی تشخیص و شناسایی مالیات دهندگان با فاکتور های ساختگی با استفاده از تکنیک های داده کاوی پرداختند این پژوهش در مجموعه مقالات پایگاه Elsevier مورد استفاده قرار گرفته است[۱۰].

در پژوهش حاضر یک دوره دو سال بین سالهای ۱۳۹۵ تا ۱۳۹۷ از اطلاعات مجرمین بانک های موجود در آگاهی ناجا با هماهنگی های بعمل آمده و با توجه به رعایت سطح طبقه بندی دریافت شد، سپس آنها را با توجه به فیلدهای مورد نیاز اعم از تاریخ وقوع جرم و ساعت وقوع و نوع و نحوه سرقت، سابقه دستگیری، محل سرقت طبقه بندی و مورد بررسی قرار گرفت. از آنجایی که یکی از مهم ترین مراحل یک روش داده کاوی، پیش پردازش داده های آن تحقیق است و پیش پردازش تعیین می کند که منجر به چه نتایجی شود و اهمیت آن به قدری است که می تواند منجر به بهترین نتیجه یا ضعیف ترین نتیجه شود. لذا در این تحقیق پیش پردازش را با توجه به مقالات و به روش کاملا اصولی انجام داده و موارد فوق را رعایت کرده ایم:

- حذف داده های نویز و پرت
- مرتب سازی داده ها
- لیبل گذاری داده ها

در نهایت اطلاعات و ج داول بدست آمده از داده های مجرمین برای خواندن نرم افزار آماده سازی و در قالب یک فایل صفحه گستر در نرم افزار Execl آماده سازی شدند. به دلیل اینکه نرم افزار بتواند داده ها را به درستی دریافت نماید این اطلاعات را به صورت عدد های ریاضی معرفی می شود. پس از پیاده سازی مرحله پیش پردازش، داده ها را در یک فایل بنام data mad

---

۲ Yee, etal
۳ Bassiouni, etal
۴ Chakravorty, etal
۵ McClendon & Meghanathan
۶ González & Velásquez

ذخیره کرده سپس در مرحله بعد برای هر نوع طبقه بندی و پیش بینی، به عنوان ورودی، وارد برنامه می شوند و با درنظر گیری تمامی پارامترها، نتایج حاصل از ارزیابی را برای دو الگوریتم بردار ماشین پشتیبان و نیز شبکه عصبی مصنوعی و درخت تصمیم ارائه کردیم و الگوریتم خوشه بندی K-means به الگوریتم های SVMو ANN و نیز الگوریتم های ترکیبی GA & SVM و GA & ANN را مورد بررسی قرار دادیم. به طور کلی میتوان گفت با بکارگیری متدولوژی مطرح شده به عنوان یک الگو، بر روی داده های جمع آوری شده از بانک اطلاعات مجرمین سرقت های مسلحانه، داده کاوی صورت پذیرد و بر اساس روش پیشنهادی، به بیان مراحل مدل پرداخته و مطابق با روشها و الگوریتم های پیشنهادی، پیاده سازی بر روی داده های جمع آوری شده و پالایش آنها انجام گرفت و به این نتیجه رسیدیم که الگوریتم شبکه عصبی کارایی بهتری نسبت به ماشین بردار پشتیبان می باشد.

لنگری زاده و اروجی در سال ۱۳۹۶، به بررسی پژوهشی با عنوان شناسه زدایی پرونده الکترونیک سلامت با استفاده از الگوریتم های یادگیری ماشین: یک مرور نظاممند پرداختند. این پژوهش مروری نظام مند بر تحقیقات اخیر می باشد، که به حذف تمامی شناسه ها از پرونده الکترونیک سلامت با استفاده از انواع روش های شناسه زدایی مبتنی بر یادگیری ماشین پرداخته. این پژوهش به صورت مروری نظام مند در بازه زمانی ۲۰۱۶ - ۲۰۰۶ در پایگاه‌های PubMed و Science direct انجام شد. مقالات با استفاده از چک لیست CASP و سپس توسط دو ارزیاب به طور مستقل بررسی و ارزشیابی شدند. در نهایت ۱۲ مقاله با معیارهای ورود مطالعه همخوانی داشتند. مقالات منتخب بر اساس روش و منابع دانش مورد استفاده، انواع شناسه ها، نوع اسناد بالینی، چالش ها و نتایج حاصل بررسی شده اند. نتایج نشان داد که در زمان انتشار داده های بالینی برای اهداف ثانویه شناسه زدایی مبتنی بر یادگیری ماشین راهکاری مناسب برای حفظ حریم خصوصی بیماران است. همچنین ترکیب الگوریتم های یادگیری ماشین و روش هایی چون تطابق الگو و عبارات منظم می تواند نیاز به داده آموزش را کاهش دهد. در پرونده های پزشکی اطلاعات شناسایی زیادی وجود دارد. این مطالعه نشان داد که روش های شناسه زدایی مبتنی بر یادگیری ماشین می توانند به طرز چشمگیری خطر افشای این اطلاعات را کاهش دهند[۵]. این نتایج با یافته های موجود در پژوهش هم راستا می باشد.

عباسی راد و مدرکی در سال ۱۳۹۶، به بررسی پژوهشی با عنوان ارائه راهکاری جهت تشخیص جرائم در داده های بزرگ با استفاده از داده کاوی پرداختند. در این تحقیق از روش خوشه بندی مبتنی بر چگالی برای کمک به تجزیه و تحلیل داده های جرم و جنایت استفاده شده. روش مورد نظر مورد تجزیه و تحلیل قرار گرفته و تمامی ابعاد آن در نظر گرفته شد و در نهایت یک روش پیشنهادی بر مبنای روش خوشه بندی مبتنی بر چگالی و با استفاده از معماری هادوپ جهت انجام خوشه بندی های مورد نظر ارائه شده. نتایج روش پیشنهادی با الگوریتم مورد استفاده پیشین (الگوریتم کامینز) مقاسیه شد و مشخص شد که روش پیشنهادی ما در پایگاه دادهایی با تعداد مختلف سرعت و عملکرد مناسب تری دارد. همچنین با توجه به ماهیت مجموعه داده ایجاد شده و روش پیشنهادی، پرس و جو مورد نظر ما با توجه به هر فیلدی که مورد نظر باشد، می تواند واقع شود و منجر به یک خوشه بندی مفهومی روی محیط گردد. بعنوان مثال در این تحقیق تعداد جرائم اتفاق افتاده در هر منطقه، خوشه بندی بر اساس جنسیت مجرمان مورد بررسی قرار گرفته است. ولی با توجه به ماهیت روش پیشنهادی می توان هر نوع خوشه بندی دیگری را نیز که مد نظر باشد تحقق بخشید. با استفاده از روش پیشنهادی و با یک محاسبه

ی ساده می توان منطقه ی دقیق جرم و محدوده ی طول و عرض جغرافیای آن را شناسایی نمود. همچنین می توان دسته بندی داده ها را بر اساس نوع جرم و یا جنسیت نیز در نظر گرفت و در صورت بروز یک جرم خاص در یک منطقه ی خاص اشخاص مورد نظر را شناسایی نمود. برخی از دستاوردهای مطالعه ما که شامل تجزیه و تحلیل الگوی جرم است می تواند به کارآگاهان در تشخیص هر چه بهتر و سریعتر جرم کمک کند، ولی جایگزین آن ها نیست. همچنین نقشه برداری داده های واقعی در داده کاوی ویژگی است که همیشه کار آسانی نیست و اغلب نیازمند داده کاو ماهر و تحلیلگر جرم با دانش خوب و همکاری نزدیک با یک کارآگاه در مراحل اولیه است[۲]. این نتایج با یافته های موجود در پژوهش هم راستا می باشد.

مانیان همکاران در سال ۱۳۹۵، به بررسی پژوهشی با عنوان طراحی الگوی داده کاوی پیشنهادی به منظور شناسایی مجرمان پرداختند. این پژوهش بر آن است تا با بهره گیری از الگوریتم های داده کاوی به تحلیل داده های ثبت شده در بانک اطلاعاتی پلیس مربوط به دستگیرشدگان توسط گشت های انتظامی تهران بزرگ در سه ماهه اول سال ۱۳۸۹ بپردازد و با استفاده از آنها، الگویی طراحی شود که به شناسایی مجرمان واقعی از بین انبوه متهمان دستگیر شده اقدام کند. این الگو می تواند به عنوان یک سامانه تصمیم یار در اختیار کارشناسان انتظامی قرار گیرد تا فرآیند شناسایی و دستگیری مجرمان واقعی با سرعت و دقت بیشتری انجام شود. این پژوهش از نوع پژوهش های داده محور بوده و بر اساس یک فرایند استاندارد داده کاوی CRISP - DM ، داده های دستگیرشدگان که شامل متغیرهای جمعیت شناختی متهمان و کلانتری محل دستگیری است، پس از یکپارچه سازی و پالایش، با استفاده از الگوریتم های CHAID , CRT ۰.۵.c و شبکه عصبی MLP مدل سازی شدند. الگوریتم C5 . 0 در فن درخت تصمیم نتایج بهتری را به لحاظ دقت شناسایی مجرمان واقعی نسبت به سایر الگوریتم های درخت تصمیم، مانند CHAID, CRT دارد؛ اما نسبت به الگوی طراحی شده توسط شبکه عصبی MLP دقت کمتری دارد. نتایج: با استفاده از الگوریتم های درخت تصمیم، در مجموع ۱۹ قانون کشف و ارائه شد. برای بررسی این قوانین، نشست خبرگان تشکیل شد و در نهایت از ۱۹ قانون استخراج شده، ۳ قانون مرتبط با موضوع مورد پژوهش شناخته شده و مورد تأیید قرار گرفت[۶]. این نتایج با یافته های موجود در پژوهش هم راستا می باشد.

ابراهیمی و همکاران در سال ۱۳۹۴، به بررسی پژوهشی با عنوان جامعیت بخشی به مجموعه داده جرائم به منظور پیشبینی و شناسایی جرائم با استفاده از تکنیک های داده کاوی پرداختند. ایده اولیه این پژوهش استفاده از تکنیک های داده کاوی در حوزه جرم شناسی به منظور رسیدن به مدل هایی جهت پیش بینی و شناسایی برخی از ویژگی های ناشناخته و مبهم جرائم است که می توان با از بین بردن بسترهای جرم زا آشکار شده توسط این مدل ها تا حدودی از وقوع رفتارهای نابهنجار و مجرمانه پیشگیری نمود. کار با ایجاد یک پیکره دادهای داخلی با مطالعه و فیش - برداری از روی تعدادی پرونده قضایی موجود در اجرای احکام شهرستان رشت آغاز گردید. بعد از پیش پردازش و داده کاوی با استفاده از الگوریتم های مختلف طبقه بندی جهت یافتن مدل بهینه تر و انجام مهندسی ویژگی روی این پیکره داده ای، ارزیابی مدل های حاصله با توجه به ویژگی هدف نوع جرم نشان داد که مدل ایجاد شده توسط الگوریتم LogitBoost دارای میانگین وزن معیار F - Measure بیشتری از مدل های دیگری است که با استفاده از الگوریتم های Bayesnet و LMT ارائه شدند. علاوه بر این بر روی اطلاعات جرائم شهر لندن تیز پروسه داده کاوی پیاده سازی گردید و نتایج ارزیابی مدل های حاصله نشان می دهند که مدل ایجاد شده توسط الگوریتم Bayesnet با توجه به ویژگی هدف نوع جرم دارای میانگین وزن معیار F- Measure بیشتری از مدل های دیگری

است که با استفاده از الگوریتم های RandomSubSpace و IBK ارائه شدند. به منظور خوشه بندی مجموعه داده داخلی از الگوریتم EM استفاده گردید که درستی عملکرد آن ۴۸/۵۷۱۴ درصد است. خوشه بندی مجموعه داده city of London Police با استفاده از الگوریتم SimplekMeans روی یازده خوشه پیاده سازی گردید که درستی عملکرد آن برابر ۲۵/۹۸۲۹ درصد است. بکارگیری الگوریتم های مختلف و محدود نساختن مدلسازی روی یک الگوریتم داده کاوی خاص را می توان به عنوان جنبه نوآورانه این مقاله در نظر گرفت. پیاده سازی جامعتر با اعمال الگوریتم های داده کاوی روی محدوده جغرافیایی و جمعیتی وسیعتر و بکارگیری روش های ترکیبی در رسیدن به مدل های منطقی تر و کامل تر در آینده می تواند مد نظر باشد. علاوه بر این سنجش عملکرد مدلها با توجه به ویژگی های هدف مختلف می تواند در دستیابی به الگوهای کارآمدتر قابل پیگیری باشد[۱]. این نتایج با یافته های موجود در پژوهش هم راستا می باشد.

یه[۷] و همکاران در سال ۲۰۱۸، به بررسی پژوهشی با عنوان شناسایی کلاهبرداری کارت اعتباری با استفاده از یادگیری ماشینی به عنوان تکنیک داده کاوی پرداختند. این پژوهش معیارهای طبقه‌بندی را با استفاده از پنج دسته بندی کننده بیزین به نام‌های بیز ساده Naïve Bayes، K2، TAN، لجستیک و J48 مورد بررسی قرار داد. ارزیابی آنها با استفاده از دو مجموعه داده انجام شد، که اولین مجموعه داده‌ها یک مجموعه داده ساختگی بود که نشان‌دهنده ویژگی‌های داده‌های کارت اعتباری و دیگری مجموعه داده تغییر یافته با استفاده از نرمال سازی داده و تکنیک‌های تحلیل مؤلفه‌های اصلی بود. به طور کلی، تمام دسته بندی کننده‌های بیزین با استفاده از داده‌های فیلتر شده، به طور قابل‌توجهی نتایج بهتری بدست آوردند. در مقایسه با نتایج بدست‌آمده از پیش‌پردازش مجموعه داده‌ها، تمامی دسته بندی کننده ها به بیش از ۹۵ ٪ صحت دست یافتند[۱۳]. این نتایج با یافته های موجود در پژوهش هم راستا می باشد.

باسیونی[۸] و همکاران در سال ۲۰۱۸، به بررسی پژوهشی با عنوان دسته بندی ایمیل های هرزنامه و HAM با استفاده از تکنیک های یادگیری ماشین پرداختند. این پژوهش مجموعه داده های متفاوت مورد استفاده برای طبقه بندی را نشان می دهد. مجموعه داده های SPAMBASE, UCI, ENRON رایج ترین مجموعه داده های مورد استفاده در طبقه بندی هستند. بیشتر کاربردها از فیلترینگ و تکنیک های یادگیری ماشین و برخی آنها را برای دست یابی به عملکرد بالا ترکیب کردند. این روش به بالاترین دقت با استفاده از روشی موثر برای طبقه بندی ایمیل های هرزنامه در مجموعه داده های UCI پایگاه هرزنامه دست پیدا می کند. این روش شامل پیش پردازش ، ILFS برای انتخاب خصوصیت و طبقه بندی داده ها با استفاده از ۱۰ طبقه بندی میباشد. ان طبقه بندی ها براساس RF ANN، رگرسیون لجستیک، ماشین بردار پشتیبان، درخت تصادفی، KNN ، جدول دقت، شبکه بیز، NB و RBF است. با توجه به نتایج به دست آمده بهترین عملکرد با استفاده از تکنیک جنگل تصادفی حاصل میشود که به دقت ۹۵.۴۵ درصد دست پیدا میکند[۸]. این نتایج با یافته های موجود در پژوهش هم راستا می باشد.

---

[۷] Yee, etal
[۸] Bassiouni, etal

چاکرا[۹] و همکاران در سال ۲۰۱۵، به بررسی پژوهشی با عنوان تکنیک های داده کاوی برای تجزیه و تحلیل داده های ساخت یافته و غیر ساخت یافته‌ی مربوط به قتل پرداختند. در این پژوهش، اثربخشی تکنیک‌های داده کاوی مانند خوشه‌بندی بر داده‌های ساختار یافته قابل دسترس برای عموم (دیوان ثبت جرایم ملی[۱۰]) در بدست آوردن اگاهی نشان داده شده. علاوه بر این، روشی برای بازیابی داده‌ها از طریق وب اسکرپینگ (استخراج داده از صفحات وب) از رسانه‌های خبری و تکنیک‌های زبان طبیعی لازم برای استخراج اطلاعات معنی‌دار ارایه شده‌است، اطلاعاتی که از طریق منابع ساختار یافته سنتی داده‌ها در دسترس نیستند. تمرکز این مقاله بر روی قتل (IPC۳۰۲)و مرتکب جنایت قتل (IPC۳۰۴، ۳۰۷ و ۳۰۸)است، که می‌تواند به جرائم اصلی دیگر تعمیم یابد. مزیت داده‌های غیر ساخت‌یافته[۱۱] دسترسی آسان به آنها است. جمع‌آوری روزانه ی چنین داده‌هایی می‌تواند برای پیدا کردن تغییرات سریع در الگوهای جنایت مورد استفاده قرار گیرد در نتیجه به پلیس و دولت هشدار می‌دهد که نسبت به هر وضعیت جدیدی که ممکن است بوجود آید واکنش نشان دهند. با این حال داده‌های ساختار یافته اگرچه بسیار دقیق‌تر هستند، در هر نیم سال منتشر می‌شوند و در نتیجه به زمان بیشتری نیاز دارند. با این حال، ایراد داده های غیر ساخت یافته فزونی و حشو در آنها است. اخبار یکسان برای چند بار تکرار می‌شود و از این رو افزونگی نیز شامل آنها میشود[۹]. این نتایج با یافته های موجود در پژوهش هم راستا می باشد.

کلندن[۱۲] و همکارش در سال ۲۰۱۵، به بررسی پژوهشی با عنوان استفاده از الگوریتم های یادگیری ماشینی برای تجزیه و تحلیل داده های جرم و جنایت پرداختند. در این تحقیق، از WEKA، یک نرم‌افزار داده کاوی متن‌باز، برای انجام یک مطالعه تطبیقی بین الگوهای جنایت خشونت‌آمیز از مجموعه داده های ارائه شده توسط ریپازیتوری Communities and Crime Unnormalized Dataset دانشگاه کالیفرنیا اروین و داده‌های آماری جرم و جنایت ایالت میسیسیپی که توسط neighborhoodscout.com فراهم شده‌است، استفاده شده. ما رگرسیون خطی، رگرسیون جمعی، و الگوریتم decision stumps را با استفاده از همان مجموعه متناهی از ویژگی‌ها، بر مجموعه داده های جوامع و جنایت (Communities and Crime) اجرا کردیم. به طور کلی، الگوریتم رگرسیون خطی بهترین الگوریتم در میان سه الگوریتم انتخاب‌شده نتیجه داد. هدف از این پروژه اثبات این مساله است که الگوریتم‌های یادگیری ماشینی که در آنالیز داده کاوی به کار می‌روند تا چه حد می‌توانند در پیش‌بینی الگوهای جنایت خشونت‌آمیز موثرباشند[۱۱]. این نتایج با یافته های موجود در پژوهش هم راستا نمی باشد.

گنزالز[۱۳] و همکارش در سال ۲۰۱۳، به بررسی پژوهشی با عنوان تشخیص و شناسایی مالیات دهندگان با فاکتور های ساختگی با استفاده از تکنیک های داده کاوی پرداختند. در این پژوهش شواهدی ارائه شده که شناسایی کاربرانی با فاکتورهای ساختگی در یک سال ممکن شود، که بسته به اطلاعات پرداختی مالیات آن‌ها، عملکرد و مشخصات تاریخی آن‌ها، با استفاده از انواع مختلف تکنیک‌های داده کاوی، این کار صورت میگیرد. ابتدا، الگوریتم‌های خوشه‌بندی مانند SOM و گاز عصبیneural gas برای

---

[۹] Chakravorty, etal
[۱۰] National Crime Records Bureau
[۱۱] Unstructured Data
[۱۲] McClendon & Meghanathan
[۱۳] González & Velásquez

شناسایی گروه‌های رفتاری مشابه در میان مالیات دهندگان استفاده می‌شوند. سپس درخت تصمیم decision tree، شبکه‌های عصبی و شبکه‌های بیزین برای شناسایی متغیرهایی که مربوط به انجام کلاهبرداری و یا عاری از کلاهبرداری هستند و برای، شناسایی الگوهای رفتار مرتبط به کار میروند و همچنین برای فهمیدن اینکه تا چه حد موارد کلاهبرداری و یا عاری از کلاهبرداری را می توان با اطلاعات موجود شناسایی کرد از این طریق به دست می آید. این کار به شناسایی الگوهای کلاهبرداری و تولید دانش کمک می‌کند که می‌تواند در امور حسابرسی انجام‌شده توسط اداره مالیات شیلی برای تشخیص این نوع از جرم مالیاتی مورد استفاده قرار گیرد[۱۰]. این نتایج با یافته های موجود در پژوهش هم راستا می باشد.

نتیجه گیری

تحلیل جرم، زیر مجموعه ای از علم جرم شناسی به شمار می رود و به معنای مطالعه بر جرایم اتفاق افتاده، شناسایی الگوها، مراحل و مشکلات و همچنین انتشار اطلاعاتی است که به پلیس برای توسعه تاکتیکها و استراتژیهای حل این الگوها و مشکلات کمک می کند. تحلیل جرم تا حد زیادی به یک مقایسه ی دقیق بین جرم فعلی و جرمهای گذشته بستگی دارد که روشهای متنوعی داشته و در تمامی این روش ها تحلیلگران الگوها و مراحل وقوع جرم را با توجه به جرایم مشابه پیشین تشخیص می دهند. پیشرفتهای موجود در تکنولوژی که امکان تحلیل تعداد زیادی داده را در مدت زمان نسبتا قابل قبولی فراهم می کند ، شالودهی مناسبی برای تحلیل جرمهای پیچیده ای همچون قتل و همچنین جرایم ساده تر ولی دارای حجم های بالا ، مانند سرقت فراهم آورده است. نمود تحلیل جرم در واقع تفکیک اعمال خلاف قانون به بخشهایی است که ماهیت و علت انجام آنها را بیان می کند. هدف تحلیل جرم یافتن اطلاعات معناداری از میان مقادیر زیادی داده، و انتشار این اطلاعات به کارآگاهان و مأمورین، برای کمک به آنها در یافتن جرایم و توقیف فعالیتهای خلاف قانون می باشد. تحلیل جرم همچنین به پیشگیری از جرم کمک شایانی می کند، زیرا جلوگیری از جرم هزینه کمتری نسبت به توقیف جرم اتفاق افتاده دارد. هیچ تردیدی وجود ندارد که جرم شناسی در آینده در میان علوم اجتماعی جایگاه والا و برجسته ای خواهد داشت .گرچه قابل انکار هم نیست که جرم شناسی آینده از جرم شناسی امروز، بدان گونه که ما آن را می شناسیم، متفاوت خواهد بود . جرم شناسی، امروزه مرحله خاصی از پیشرفت را طی کرده و در حال رشد و بلوغ خود می باشد. تطابق جرایم از آن جهت برای پلیس مهم است که با تحقق این هدف پلیس قادر است شک ها و نظریه های احتمالی را شناسایی کند . در واقع فرآیند تطبیق دادن جرم کشف شده ای که به تازگی رخ داده است با جرایم مشابه پیشین ،جهت مشخص شدن مجهولات جرم فعلی کشف شده را "تطبیق جرم" گویند. این واقعیت اثبات شده است که نه تنها مکان وقوع جرم و اطلاعات جغرافیایی مربوط به آن هرگز برای دستیابی به الگو و مشخص کردن تمرکز و تراکم جرم کافی نمی باشد بلکه متغیرهای غیرفضایی وغیر جغرافیایی برای تحقق اهداف فوق ضروری به نظر می رسند. از این رو ترکیب داده های محلی و دادههای رفتاری مجرم بسیار کارآمدتر به نظر می رسد. از مهمترین چالشها و محدودیتهای انجام این تحقیق به عدم وجود مقالات و متدولوژی های مطرح شده بصورت جامع و مرتبط با بررسی مجرمین پلیس کشورمان می توان اشاره کرد. محرمانگی اطلاعات مجرمین یکی دیگر از مشکلات سر راه این تحقیق بود.

در نهایت بطور خلاصه مهمترین نتایج به دست آمده از این تحقیق عبارتند از:

- داده کاوی نرم که یکی از مراحل آن خوشه بندی نرم مورد استفاده قرار گرفت می تواند

برای سرعت انجام کار در بانک های مجرمین بزرگ مورد استفاده قرار گیرد.

- داده کاوی سخت که یکی از مراحل آن خوشه بندی با دقت بالا می تواند یکی از الگوریتم

های تاثیر گذار برای آماده سازی داده ها برای پیش بینی مورد استفاده قرار گیرد تا سطح کارایی سیستم را ارتقاء دهد .

- چنانچه این تحقیق بصورت عملی در آگاهی ناجا و با اطلاعات کاملتر و جامع تر مورد استفاده قرار گیرد، شاهد تاثیر آن بر افزایش کشف جرم خواهیم بود.

- در این تحقیق استفاده از الگوریتم ژنتیک می تواند تاثیر بسزایی در شناسایی مجرمینی که شباهت بیشتری با عمل مجرمانه دارند کمک نماید .

منابع

[۱]. ابراهیمی، مجیب، روشندل، ابوالقاسم، آقایی، جان احمد، ۱۳۹۴، جامعیت بخشی به مجموعه داده جرائم به منظور پیشبینی و شناسایی جرائم با استفاده از تکنیک های داده کاوی، فصلنامه صنایع الکترونیک، دوره ۶، شماره ۴، زمستان ۱۳۹۴.

[۲]. عباسی راد، عبدالرضا، مدرکی، عباس، ۱۳۹۶، ارائه راهکاری جهت تشخیص جرائم در داده های بزرگ با استفاده از داده کاوی، نشریه فناوری و پژوهش های نوین، دوره ۱، شماره ۱، پاییز و زمستان ۱۳۹۶.

[۳]. علی اکبر نیک نفس، حمید میروزیری، عارف طهماسب. ۱۳۹۳. بهبود کلاس بندی داده‌های نامتوازن با استفاده از الگوریتم‌های یادگیری ماشین .وزارت علوم، تحقیقات و فناوری - دانشگاه شهید باهنر کرمان - دانشکده فنی.

[۴]. کوشا، نوشین، ۱۳۹۵، بهینه سازی Query ها در sql server، کنفرانس بین المللی مهندسی کامپیوتر و فناوری اطلاعات، تهران.

[۵]. لنگری زاده، مصطفی، اروجی، اعظم، ۱۳۹۶، شناسه زدایی پرونده الکترونیک سلامت با استفاده از الگوریتم های یادگیری ماشین: یک مرور نظاممند، مجله انفورماتیک سلامت و زیست پزشکی، مرکز تحقیقات انفورمانیک پزشکی، دوره چهارم، شماره دوم، ص ۱۵۴-۱۶۷، ۱۳۹۶.

[۶]. مانیان، امیر، جمالو، محمد، بیدل، معصومه، ۱۳۹۵، شناسه زدایی پرونده الکترونیک سلامت با استفاده از الگوریتم های یادگیری ماشین: یک مرور نظاممند، دانشگاه آزاد اسلامی، واحد علوم تحقیقات تهران.

[۷]. یوسف وند، زهرا . السادات نقیبیان، سیمین ، ۱۳۹۵، تشخیص حالت چهره با استفاده از تصویر چهره و بکارگیری تبدیل موجک و شبکه عصبی فازی، کنفرانس بین المللی مهندسی کامپیوتر و فناوری اطلاعات.

[8]. Bassiouni, M., Ali, M., & El-Dahshan, E. A. (2018). Ham and Spam E-Mails Classification Using Machine Learning Techniques. Journal of Applied Security Research, 13(3), 315-331.
[9]. Chakravorty, S., Daripa, S., Saha, U., Bose, S., Goswami, S., & Mitra, S. (2015). Data mining techniques for analyzing murder related structured and unstructured data. American Journal of Advanced Computing, 2(2), 47-54.
[10]. González, P. C., & Velásquez, J. D. (2013). Characterization and detection of taxpayers with false invoices using data mining techniques. Expert Systems with Applications, 40(5), 1427-1436.
[11]. McClendon, L., & Meghanathan, N. (2015). Using machine learning algorithms to analyze crime data. Machine Learning and Applications: An International Journal (MLAIJ), 2(1), 1-12.
[12]. Varma, M. C., & Kumar, B. S. S. (2018). Detection of Intrusion Using Decision Tree Based Data Mining.
[13]. Yee, O. S., Sagadevan, S., & Malim, N. H. A. H. (2018). Credit Card Fraud Detection Using Machine Learning As Data Mining Technique. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 10(1-4), 23-27.