جامعیت بخشی به مجموعه داده جرائم به منظور پیش‌بینی و شناسایی جرائم با استفاده از تکنیک‌های داده کاوی

مجیب ابراهیمی[1]، سید ابوالقاسم میر روشندل[2]، جان احمد آقایی[3]

۱ کارشناس ارشد مهندسی کامپیوتر- نرم‌افزار، دانشگاه علوم و تحقیقات گیلان

۲ استادیار گروه مهندسی کامپیوتر، دانشکده فنی، دانشگاه گیلان، mirroshandel@guilan.ac.ir

۳ مربی گروه حقوق جزاء، دانشگاه آزاد اسلامی واحد انزلی

تاریخ دریافت: ۹۳/۵/۳۱ تاریخ پذیرش: ۹۴/۵/۱۷

چکیده

رویارویی انسان با حوادث شرورانه و مجرمانه در زندگی اجتماعی غیرقابل اجتناب بوده و انسان همواره نیازمند شناخت محیط زیست خود است. به کارگیری شیوه‌های نظام‌مند جهت شناسایی، کشف و پیش‌گیری از وقوع جرائم در تحلیل جامعه برای رسیدن به یک سیستم تحلیل جرم رو به گسترش است. با توجه به گسترش اطلاعات و توسعه سیستم‌های اطلاعاتی در سازمان-ها در این مقاله با استفاده از روش‌های داده‌کاوی به تحلیل و بررسی اطلاعات گردآوری شده جرائم و بانک‌های اطلاعاتی موجود پرداخته شده است. استفاده از ابزارها و الگوریتم‌های داده‌کاوی ماهیت پیچیده داده‌های بزه بزه‌کاری و روابط نامحسوس میان داده‌ها را مدل کرده و الگوهای جرم را شناسایی، کشف و در سدد پیش‌گیری برمی‌آید. از الگوریتم‌های طبقه-بندی مدل‌های بهینه برای پیش‌بینی ویژگی‌های جرائم ارتکابی آینده و از الگوریتم‌های خوشه‌بندی در شناسایی نوع جرم روی مجموعه داده‌های گردآوری شده استفاده گردید.

کلیدواژه

داده‌کاوی، طبقه‌بندی، خوشه‌بندی، جرم‌شناسی، مدلسازی.

مقدمه

امروزه به دلیل رشد اطلاعات، کاربرد کامپیوتر در زندگی بشر ابعاد گسترده‌ای پیدا کرده است. در بخش‌هایی که حتی روزی فکرش هم به ذهن خطور نمی‌کرد، امروزه تحلیل و محاسبات بدون استفاده از روش‌های کامپیوتری امکان‌پذیر نیست. از اینگونه موارد می‌توان به مسئله جرم‌شناسی و کشف جرم اشاره نمود. پارامترهای بسیار متنوع و گوناگون دخیل در بحث تحلیل جرم و جنایت، استفاده از تکنیک‌های مطرح در علوم مختلف را می‌طلبد. ویژگی-های بزه‌کار، نحوه انجام عمل مجرمانه، ویژگی‌های بزه‌دیده و رابطه بین بزه‌کار و بزه‌دیده همگی در مطالعه یک عمل مجرمانه و شناخت آن موثر هستند. اما چگونه رابطه بین این پارامترها قابل تشخیص است؟ چگونه می‌توان از اطلاعات موجود جرائم ارتکابی مدلی برای پیش‌بینی و به دنبال آن پیشگیری ارائه نمود؟ اطلاعات گردآوری شده از ویژگی‌های جرم و جنایت در یک پایگاه‌داده به مانند یک معدن طلا ارزشمند هستند. کشف دانش پنهان در این پایگاه‌های داده کلید حل مسئله است. روش‌های و تکنیک‌های داده‌کاوی به عنوان ابزارهای کاوش مخازن داده به استخراج دانش نهفته در دل آنها می‌پردازد. در حال حاضر مطالعات خوبی در امر داده‌کاوی اطلاعات اعمال مجرمانه در دنیا انجام شده است اما متاسفانه در داخل کشور فعالیت چندانی به چشم نمی‌خورد. استفاده از روش‌های داده‌کاوی در شناسایی، پیش‌بینی و پیشگیری جرم و جنایت در کشور می‌نواند ثمرات نوآورانه‌ای به همراه داشته باشد. در این مقاله سعی شده در ابتدا با بیان مطالعات مرتبط و کارهای انجام شده در این زمینه اهمیت موضوع را روشن سازد تا زمینه‌ساز حرکت‌های جدی در این حوزه شود. سپس به صورت موردی در بخش پنج روی اطلاعات جرائم شهرستان رشت که با مطالعه‌ی پرونده‌های جنایی افراد گردآوری شد و همچنین اطلاعات جرائم شهر لندن به صورت مجزا با استفاده از الگوریتم-های داده‌کاوی مختلف برای رسیدن به مدلهایی مطلوب‌تر جهت پیش‌بینی و شناسایی جرائم کار شد که در نهایت مدل‌هایی که با استفاده از تکنیک‌ها و الگوریتم‌های داده‌کاوی Bayesnet، LogitBoost، LMT، IBK، RandomSubSpace، EM و SimpleKMeans بدست آمد به عنوان مدل‌های بهینه گزارش گردید.

کارهای پیشین

توانایی پیش‌بینی زمان، مکان و یا نوع جرم بعدی یا مجموعه جرائمی که در آینده رخ خواهند داد یک مفهوم جامع بوده که در حال حاضر امکان‌پذیر نیست. البته تلاش‌های بسیاری در عرصه پیش‌بینی جرائم انجام شده که هر یک از آنها موفقیت‌های محدودی داشتند. بسیاری از تلاش‌های صورت گرفته مربوط به یافتن ارتباط بین جرائم با مجرمین یا یک نوع جرم معین است. تکنیک‌های تحلیل و پیش‌بینی جرم و جنایت که در طول زمان پالایش شده و موفقیت‌های محدودی را در زمینه‌های مختلف بدست آورده در سه دسته قابل تمرکز هستند: ۱) تکنیک‌های سیستم اطلاعات جغرافیایی (GIS)، ۲) روش‌های آماری، ۳) تکنیک‌های کشف دانش و داده‌کاوی [۲]. تشخیص، پیش‌بینی و پیشگیری از وقوع جرائم با داده‌کاوی یک ایده جدید و هیجان‌انگیز است که به وسیله روش‌های آماری، یادگیری ماشین، هوش مصنوعی، جرم‌شناسی، روان‌شناسی و فناوری‌های پایگاه داده به ارمغان می‌آید. در تحقیقی توسعه ابزارهای پژوهشی که از قدرت محاسباتی به عنوان یک مکانیزم برای کمک به حل مسائل عظیم و حجیم جرم و جنایت بهترین استفاده را می‌کنند و نیازمند استراتژی‌های مختلفی برای تحقیقات هستند مصور ساخته شده است [۳]. تکنیک‌های خوشه‌بندی داده‌ها را براساس شباهتشان در یک کلاس قرار می‌دهند از این رو می‌توان مظنونانی که دارای حالت و ویژگی‌های مشابه هستند شناسایی نمود یا نوع جنایت ارتکابی را از میان گروه‌های مختلف جرائم تشخیص داد. به منظور شناسایی و گروه‌بندی انواع جرائم مدلی براساس تکنیک‌های خوشه‌بندی توسط کارلیس[۱] و همکارانش ارائه گردید که یک مدل ترکیبی پواسون چند متغییره محدود با ساختار کواریانسی دو طرفه بود [۴].

قوانین انجمنی الگوهای مکرر موجود در داده‌ها هستند که می‌توانند هرگونه اختلاف را به عنوان یک نفوذ تشخیص دهند. برای اولین بار از تکنیک‌های کشف قوانین انجمنی فازی[۲] توسط بوکزاک و همکارانش در تحلیل داده‌های جنایی استفاده شد [۵]. استخراج قوانین انجمنی فازی در مطالعه جرم و جنایت بسیار مفید ارزیابی گردید. هزاران قانون کشف شده اولیه نیاز به غربال کردن در جهت پیدا کردن الگوهای جالب و معنی‌دار توسط پرسنل اجرای احکام دارند. نتایج نهایی بدست آمده نشان‌دهنده سازگاری الگوهای کشف شده جرم و جنایت در سطوح مختلف جامعه است. طبقه‌بندی اغلب برای پیش‌بینی روند جرم و جنایت استفاده می‌شود که زمان شناسایی اشخاص بزهکار را کاهش می‌دهد. این امر نیاز به آموزش و بررسی کامل پایگاه داده دارد تا با حداقل رسانی مقادیر گمشده بتواند دقت پیش‌بینی را بهبود دهد. جرائم کامپیوتری به یک مسئله جهانی تبدیل شده است. مطالعه‌ای که با استفاده از رگرسیون روی اطلاعات مربوط به استفاده از اینترنت برای پیش‌بینی جرائم رایانه‌ای صورت گرفته، دو عامل میزان استفاده از کامپیوتر و عضویت در شبکه‌های اجتماعی را به عنوان متغییرهای اصلی پیش‌بینی کننده میزان جرائم کامپیوتری معرفی کرده است. این عوامل فرصتی را به وجود می‌آورند که جوانان با بحث و گفتگو و تبادل نظر در فضای مجازی دانش خود را افزایش داده و نحوه انجام دانلودهای غیر قانونی و همچنین بدست آوردن شناسه‌های شخصی افراد را یاد بگیرند. علاوه بر این مشخص گردید که میزان جرائم کامپیوتری در مردان بیشتر از زنان و با افزایش تحصیلات دانشگاهی و کسب مهارت‌های کامپیوتری احتمال اینگونه جرائم در افراد افزایش می‌یابد [۶].

به طور کلی کاربرد تکنیک‌های داده‌کاوی در شناخت جرائم را می‌توان در قالب دو دسته اقدامات در نظر گرفت. اولین دسته شامل اقداماتی می‌شوند که قبل از وقوع جرائم به منظور پیش‌بینی و پیشگیری از ارتکاب جرم انجام می‌گیرند و دسته دوم پیرامون اقدامات انجام شده بعد از وقوع جرم به منظور بررسی و کشف مدارک و شواهد جرم پس از وقوع آن است [۷]. به طور کلی در این چارچوب می‌توان یک دسته‌بندی کاربردی از کارهای انجام‌شده در زمینه شناسایی جرائم، پیش‌بینی جرائم و پیشگیری جرائم به تفکیک کاربرد تکنیک‌های داده‌کاوی در این موارد داشت که در جدول ۱ قابل مشاهده است. اما هیچ تفاهمی در مورد چگونگی کنترل مردم به عنوان عاملان جرائم و منبع اطلاعاتی پژوهش‌ها وجود ندارد. همچنین انتخاب روش مناسب براساس نوع جرم و جنایت صورت می گیرد و هیچ اجماعی در مورد استفاده از یک روش خاص مدنظر نیست. در نهایت تفسیر نتایج به نظر آمارشناسان، جامعه‌شناسان و محققان جرم وجنایت بستگی دارد.

---

۱ - Karlis

۲ - Fuzzy Association Rules

جدول ۱. چارچوب کاربرد تکنیک‌های داده‌کاوی در مدل‌سازی جرم و جنایت

حوزه‌های کاربرد	تکنیک‌های مورد استفاده	مرجع
شناسایی جرائم	خوشه‌بندی	کارلیس و همکاران [۴]
	خوشه‌بندی	آدلری و همکاران [۳]
	خوشه‌بندی	مورتاق و همکاران [۸]
	خوشه‌بندی باینری	ماند و همکاران [۹]
پیش‌بینی جرائم	پیش‌بینی- رگرسیون	کراپسیوگلو و همکاران [۱۰]
	پیش‌بینی – رگرسیون	مون و همکاران [۱۱]
	پیش‌بینی	لیو و همکاران [۱۲]
	پیش‌بینی-رگرسیون	دالسیو و همکاران [۱۳]
	پیش‌بینی مبتنی بر نقاط جرم خیز	لیو و همکاران [۱۱]
	پیش‌بینی – سری‌های زمانی	دیدمن [۱۳]
	پیش‌بینی – رگرسیون	فریلیچ و همکاران [۱۴]
	خوشه‌بندی – پیش‌بینی	ایکس‌یوای و همکاران [۱۵]
	خوشه‌بندی – پیش‌بینی	هادجیدی [۱۶]
	خوشه‌بندی – قوانین انجمنی	مالاتی و همکاران [۱۷]
	قوانین انجمنی فازی	بوکزاک و همکاران [۵]
پیش‌گیری جرائم	فازی سام	لی و همکاران [۱۸]
	ترکیبی از تکنیک‌های رگرسیون، شبکه عصبی و شبکه بیزین	اُتلی و همکاران [۱۹]
	رگرسیون لجستیک	دالسیو و همکارانش [۲۰]

پیکره‌های داده‌ای پیشنهادی

دستیابی به یک نتیجه و مدل مطلوب و کارآمد در کشف دانش و داده‌کاوی نیازمند دسترسی به پایگاه داده‌ها و مجموعه داده‌های معتبر است. امروزه پیشرفت‌های زیادی در ابزارهای گردآوری اطلاعات به صورت کامپیوتری فراهم گردیده است. اطلاعات جمع-آوری شده ویژگی‌هایی از محیط مورد مطالعه در اختیار افراد قرار می‌دهد که هر چقدر دقیق‌تر و جامع‌تر باشد می‌توان ارزیابی کارامدتری از آن محیط داشت. اطلاعات مربوط به جرم وجنایت ماهیتاً بدلیل داشتن پارامترهای مختلفی از ویژگی‌های جمعیت-شناختی، جغرافیایی، اجتماعی، فرهنگی و حاکمیتی بسیار پیچیده هستند که در هر مطالعه باید مورد توجه قرار گیرند. مجموعه داده پیشنهادی شامل اطلاعاتی از جرائم سال‌های ۸۹، ۹۰ و ۹۱ در سطح شهر رشت و حومه آن است که از روی پرونده‌های قضایی مجرمان موجود در اجرای احکام شهرستان رشت فیش‌برداری و به صورت بک پیکره داده‌ای منحصر به فرد در آمد. شهرستان رشت با مساحت ۱۳۷ کیلومتر مربع و جمعیتی بالغ بر ۵۱۹۴۸۱ (برآورد سال ۸۴) مرکز استان گیلان و یکی از کلان شهرهای ایران محسوب می‌گردد [۲۱]. محدودیت‌های زمانی، دسترسی و اطلاعاتی اعمال شده موجب گردید که در کار جمع‌آوری اطلاعات به نمونه‌گیری محدود کفایت شود. مسلم است که در اختیار داشتن یک نمونه آماری جامع در تحلیل درست و رسیدن به یک مدل واقعی‌تر از محیط بسیار تاثیر گذار است. مجموعه داده گردآوری شده دارای ویژگی‌های زیر است

● نوع جرم	● موقعیت جغرافیایی محل وقوع جرم	● محل تولد مجرم	
● سن مجرمان	● نوع ارتباط مجرم باهمدستانش	●جنسیت مجرم	
● نام شهر یا روستای محل زندگی مجرم	● شهرنشین یا روستایی بودن مجرم	● وضعیت بومی بودن مجرم	
● میزان سواد مجرم	● وضعیت مصرف مواد مخدر و مشروبات الکلی	● سال وقوع جرم	
● ماه وقوع جرم	● روز وقوع جرم	● ساعت وقوع جرم	● شغل مجرم
● وضعیت تاهل مجرم	● وضعیت سوء پیشینه مجرم	● شریک داشتن مجرم	● تعداد شرکاء
● سن همدستان مجرم	● جنسیت همدستان مجرم	● احکام صادره برای مجرمان	● وضعیت مالی مجرم
● جنسیت بزه‌دیده	● سن بزه‌دید	● محل تولد بزه‌دیده	● سطح سواد بزه‌دیده
● شغل اصلی بزه‌دیده	● شهرنشین یا روستایی بودن بزه‌دیده	● وضعیت بومی بودن بزه‌دیده	
● نام شهر یا روستای محل زندگی بزه‌دیده	● نوع رابطه بین بزه‌کاران و بزه‌دیدگان	● وضعیت مالی بزه‌دیده	
● وضعیت تاهل بزه‌دیده			

علاوه بر این از یک پیکره داده‌ای خارجی تهیه شده از یک مجموعه داده خارجی نیز استفاده گردید. این مجموعه داده با عنوان City Of London شامل اطلاعاتی از ویژگی‌های جرائم خیابانی شهر لندن و حومه آن طی سال‌های ۲۰۱۱ و ۲۰۱۲ است که به پلیس شهر لندن گزارش شدند [۲۲]. در تهیه این مجموعه داده مشورت‌های شدید دراز مدتی با دفتر کمیساریای اطلاعات و متخصصان حفاظت اطلاعات در وزارت کشور به منظور حفظ حریم خصوصی افراد صورت گرفته تا خطرات احتمالی به حداقل برسد در حالی که هنوز هم بتوان به اهداف روشن و مفیدی از مجرمان دست یافت. داده‌های نیروی پلیس در دفترخانه و وزارت دادگستری از طریق یک فرایند کنترل کیفیت دقیق شامل اعتبارسنجی فرمت، تست خودکار، تایید دستی و تایید توسط دو فرد خبره به صورت مجزا انجام شد. پیکره داده‌ای City Of London شامل شش ویژگی زیر است:

- سال وقوع جرم
- ماه وقوع جرم
- نوع جرم
- موقعیت شهری محل وقوع جرم
- مرزهای برداری دیجیتال
- مرجع گزارشات و حوزه استحفاظی محل وقوع جرم

پیش‌پردازش

در مورد پیکره داده‌ای داخلی ابتداً یک پیش‌پردازش اولیه به صورت دستی روی ویژگی‌ها به منظور بالا بردن کیفیت فهم صورت گرفت. در ادامه فایل CSV مجموعه داده مورد نظر با نرم‌افزار Weka 3.7 باز گردید. از فیلتر Numeric to Nominal [۳] برای تبدیل ویژگی‌هایی مثل .Year که مقادیر عددی دارند و به صورت یک ویژگی عددی شناخته می‌شوند به یک متغییر اسمی استفاده می‌شود. در مرحله پیش‌پردازش مشخصات آماری مثل وزن و فراوانی ارزش‌های هر ویژگی مشخص می‌گردد. به همین صورت روی پیکره داده‌ای City Of London Police نیز عملیات پیش‌پردازش انجام شد.

داده‌کاوی و ارزیابی

در مرحله داده‌کاوی بعد از اعمال الگوریتم‌های مختلف طبقه‌بندی و خوشه‌بندی روی مجموعه داده‌های پیشنهادی با استفاده الگوریتم‌هایی که عملکرد بهتری داشتند، مدل‌هایی در جهت شناخت الگوها و تحلیل مجموعه داده مورد مطالعه ایجاد گردید. از الگوریتم‌های طبقه‌بندی Bayesnet، LogitBoost ، LMT، IBK و RandomSubSpace در ارائه مدل‌هایی به منظور پیش‌بینی جرائم و از الگوریتم‌های EM وSimpleKMeans در تهیه مدل‌هایی به منظور شناسایی جرائم استفاده گردید. مدل یادگیر ایجاد شده به

۳ - عددی به اسمی

دانشی می‌انجامد که باید در مرحله ارزیابی مورد تحلیل قرار گیرد تا میزان صحت و کارایی دانش و الگوریتم یادگیرنده مدل مشخص شود. معیارها و روش‌های مختلفی برای سنجش مدل یادگیری وجود دارد. در اینجا از روش ارزیابی ۱۰ دسته‌ای متقابل[۴] برای ارزیابی مدل‌های بدست آمده استفاده گردید. از معیار فراخوانی[۵] و معیار دقت[۶] برای بررسی الگوریتم‌های مذکور استفاده گردید همچنین معیار F–Measure یک معیار ترکیبی از معیارهای فراخوانی و دقت است که در مواردی که نتوان برای هر دوی این معیارها نسبت به یکدیگر اهمیت ویژه‌ای قائل شد، مورد استفاده قرار می‌گیرد [۲۲].

جدول ۲. تاثیر حذف ویژگی‌های مشخص روی درصد درستی

ویژگی حذف شده	درصد درستی
موقعیت جغرافیایی محل وقوع جرم	۷۶,۶۶۶۷
شهرنشین یا روستایی بودن مجرم	۷۸,۰۹۵۲
سن مجرمان	۷۸,۰۹۵۲
نام شهر یا روستای محل زندگی مجرم	۷۷,۶۱۹
محل تولد مجرم	۷۹,۰۹۷۶
جنسیت مجرم	۷۹,۰۴۷۶
وضعیت بومی بودن مجرم	۷۸,۰۹۵۲
میزان سواد مجرم	۷۸,۵۷۱۴
سال وقوع جرم	۷۷,۶۱۹
ماه وقوع جرم	۷۷,۱۴۲۹
روز وقوع جرم	۷۸,۵۷۱۴
ساعت وقوع جرم	۷۸,۵۷۱۴
شغل مجرم	۷۵,۷۱۴۳
وضعیت تاهل مجرم	۷۵,۲۳۸۱
وضعیت سوء پیشینه مجرم	۷۷,۱۴۲۹
شریک داشتن مجرم	۷۸,۰۹۵۲
تعداد شرکاء	۷۸,۰۹۵۲
نوع ارتباط مجرم با همدستانش	۷۶,۶۶۶۷
جنسیت همدستان مجرم	۷۸,۰۹۵۲
سن همدستان مجرم	۷۸,۰۹۵۲
احکام صادره برای مجرمان	۷۸,۰۹۵۲
وضعیت مالی مجرم	۷۶,۱۹۰۵
وضعیت مصرف مواد مخدر و مشروبات الکلی	۷۷,۶۱۹
جنسیت بزه‌دیده	۷۶,۶۶۶۷
سن بزه‌دیده	۷۳,۸۰۹۵
محل تولد بزه‌دیده	۷۸,۰۹۵۲
شهرنشین یا روستایی بودن بزه‌دیده	۷۸,۰۹۵۲
نام شهر یا روستای محل زندگی بزه‌دیده	۷۸,۵۷۱۴
وضعیت بومی بودن بزه‌دیده	۷۸,۰۹۵۲
سطح سواد بزه‌دیده	۷۶,۱۹۰۵
شغل اصلی بزه‌دیده	۸۰
وضعیت تاهل بزه‌دیده	۷۸,۰۹۵۲
وضعیت مالی بزه‌دیده	۷۵,۲۳۸۱
نوع رابطه بین بزه‌کاران و بزه‌دیدگان	۷۴,۲۸۵۷

داده‌کاوی و ارزیابی مجموعه داده داخلی

با توجه به بررسی‌های صورت گرفته به کمک الگوریتم‌های طبقه‌بندی، استفاده از الگوریتم LogitBoost مطلوب‌ترین مدل برای پیش‌بینی جرائم و مدلی با استفاده از الگوریتم EM برای شناسایی جرائم از روی مجموعه داده داخلی و با ویژگی هدف نوع جرم[۷] ایجاد نمود. برای رسیدن به مدل بهینه‌تر که درصد درستی بالاتری داشته باشد با توجه به الگوریتم LogitBoost مهندسی ویژگی انجام گردید تا تاثیر حذف ویژگی‌های مختلف روی میزان کارایی مدل مشخص گردد (جدول ۲).

همانگونه که در جدول ۲ قابل مشاهده است حذف هر یک از سه ویژگی محل تولد مجرم، جنسیت مجرم و شغل مجرم بالاترین تاثیر را در افزایش درصد درستی داشته است. در ادامه برای بالا بردن درصد درستی مدل‌ها این سه ویژگی از مجموعه داده اولیه حذف و مجموعه داده هدف ایجاد می‌گردد. معیار F–Measure مدل‌های ایجاد شده با استفاده از سه الگوریتم طبقه‌بندی Bayesnet، LogitBoost و LMT روی مجموعه داده هدف که بهینه‌تر به نظر می‌آمدند، نشان می‌دهند که مدل ایجاد شده با استفاده از الگوریتم Bayesnet کلاس رابطه نامشروع را با دقت بالاتری از دو مدل دیگر پیش‌بینی می‌نماید و مدل ایجاد شده با استفاده از الگوریتم LogitBoost کلاس‌های سرقت، ضرب و جرح

۴ -10 -Fold-Cross Validation
۵ - Recall
۶ - Precision
۷ - Crime_Type

عمدی، فحاشی و مشروبات الکلی دارای بالاترین میزان دقت است و همچنین مدل ایجاد شده با استفاده از الگوریتم LMT نیز کلاس ضرب و جرح عمدی را با دقتی برابر با الگوریتم LogitBoost پیش‌بینی می‌نماید. درصد درستی مدل‌های طبقه‌بندی و خوشه-بندی با استفاده از الگوریتم‌های مذکور روی ویژگی هدف نوع جرم مطابق جدول ۳ است.

جدول۳. درصد درستی هر الگوریتم روی مجموعه داده داخلی

Crime_Type	الگوریتم	
۷۶,۱۹۰۵	Bayesnet	طبقه بندی
۸۰,۴۷۶۲	LogitBoost	
۷۵,۲۳۸۱	LMT	
۴۸,۵۷۱۴	EM	خوشه بندی

از روی اطلاعات جدول فوق مشاهده می‌شود که مدل ایجاد شده با استفاده از الگوریتم LogitBoost در بخش پیش‌بینی درصد درستی بالاتری دارد و حذف ویژگی‌های کم تاثیرتر باعث بهبود عملکرد مدل گردیده است.

داده‌کاوی و ارزیابی مجموعه داده City Of London Police

با اعمال الگوریتم‌های طبقه‌بندی مختلف به منظور رسیدن به مدل‌هایی برای پیش‌بینی جرائم مشخص گردید که استفاده از الگوریتم Bayesnet دارای بالاترین میزان کارایی روی این مجموعه داده است و همچنین به کمک الگوریتم خوشه‌بندی SimpleKMeans مدلی جهت شناسایی جرائم ارائه گردید. برای دستیابی به مدل‌هایی با کارایی بالاتر در بخش پیش‌بینی جرائم، مهندسی ویژگی صورت گرفت و با حذف هر یک از ویژگی‌های مجموعه داده City Of London Police تاثیر آن روی درصد درستی مدل‌های حاصله با استفاده از الگوریتم Bayesnet مشخص گردید (جدول ۴).

جدول ۴. تاثیر حذف ویژگی‌های مشخص روی درصد درستی

ویژگی حذف شده	سال وقوع جرم	ماه وقوع جرم	مرجع گزارشات و حوزه استحفاظی	موقعیت شهری و محل وقوع جرم	مرزهای برداری دیجیتال
درصد درستی	۴۹,۴۴۳	۴۹,۱۲۰۴	۵۲,۵۸۱۸	۴۵,۱۴۴۵	۵۲,۷۴۰۶

نتایج حاصله در جدول ۴ نشان می‌دهد که حذف ویژگی‌ها تاثیر مثبتی در افزایش درصد درستی مدل‌ها نداشته است، بنابراین مجموعه داده هدف بدون تغییر در مجموعه داده اولیه، تثبیت می-گردد. کارایی مدل‌های ایجاد شده برای پیش‌بینی ممکن است با توجه به ویژگی‌های هدف مشخص روی یک کلاس معین متفاوت باشند. به عنوان مثال با توجه به معیار F–Measure روی ویژگی هدف نوع جرم مدل ایجاد شده با استفاده از الگوریتم Bayesnet کلاس سایر انواع سرقت‌ها، دزدی از مغازه، جرائم تخریب و آتش‌سوزی عمدی، سایر جرائم را با دقت بالاتری پیش‌بینی می‌نماید و همچنین کلاس جرائم وسایل نقلیه، جرم و جنایت خشونت آمیز، سرقت از منزل، مواد مخدر و اخلال در نظم عمومی و حمل اسلحه در مدل ایجاد شده توسط الگوریتم IBK و کلاس رفتارهای ضد اجتماعی در مدل ایجاد شده توسط الگوریتم RandomSubSpace کارایی بالاتری را نشان می‌دهند. درصد درستی مدل‌های طبقه‌بندی و خوشه‌بندی با استفاده از الگوریتم‌های مذکور و با توجه به ویژگی هدف نوع جرم مطابق جدول ۵ است.

جدول ۵. درصد درستی هر الگوریتم روی مجموعه داده خارجی

Crime Type	الگوریتم	
۵۴,۲۲۶۷	Bayesnet	طبقه‌بندی
۵۲,۳۳۴۱	IBK	
۵۲,۳۳۴۱	RandomSubSpace	
۲۵,۹۸۲۹	SimpleKMeans	خوشه‌بندی

از روی اطلاعات جدول ۵ مشاهده می‌شود که الگوریتم Bayesnet در بخش پیش‌بینی درصد درستی بیشتری را نشان می‌دهد.

نتیجه‌گیری

ایده اولیه این مقاله استفاده از تکنیک‌های داده‌کاوی در حوزه جرم‌شناسی به منظور رسیدن به مدل‌هایی جهت پیش‌بینی و شناسایی برخی از ویژگی‌های ناشناخته و مبهم جرائم است که می‌توان با از بین بردن بسترهای جرم زا آشکار شده توسط این مدل‌ها 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 پیش‌پردازش و داده‌کاوی با استفاده از الگوریتم‌های مختلف طبقه بندی جهت یافتن مدل بهینه‌تر و انجام مهندسی ویژگی روی این پیکره داده‌ای، ارزیابی مدل‌های 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 مدل‌های دیگری است که با استفاده از الگوریتم‌های Bayesnet و LMT ارائه شدند. علاوه بر این بر روی اطلاعات جرائم شهر لندن نیز پروسه داده‌کاوی پیاده‌سازی گردید و نتایج ارزیابی مدل‌های حاصله نشان می‌دهند که مدل ایجاد شده توسط الگوریتم Bayesnet با توجه به ویژگی هدف نوع جرم دارای میانگین وزن معیار F-Measure بیشتری از مدل‌های دیگری است که با استفاده از الگوریتم‌های RandomSubSpace و IBK ارائه شدند. به منظور خوشه‌بندی مجموعه داده داخلی از الگوریتم EM استفاده گردید که درستی عملکرد آن ۴۸/۵۷۱۴ درصد است. خوشه‌بندی مجموعه داده city of London Police با استفاده از الگوریتم SimpleKMeans روی یازده خوشه پیاده‌سازی گردید که درستی عملکرد آن برابر ۲۵/۹۸۲۹ درصد است. بکارگیری الگوریتم‌های مختلف و محدود نساختن مدلسازی روی یک الگوریتم داده‌کاوی خاص را می‌توان به عنوان جنبه نوآورانه این مقاله در نظر گرفت. پیاده‌سازی جامعتر با اعمال الگوریتم‌های داده‌کاوی روی محدوده جغرافیایی و جمعیتی وسیعتر و بکارگیری روش‌های ترکیبی در رسیدن به مدل‌های منطقی‌تر و کامل‌تر در آینده می‌تواند مدنظر باشد. علاوه بر این سنجش عملکرد مدل‌ها با توجه به ویژگی‌های هدف مختلف می‌تواند در دستیابی به الگوهای کارآمدتر قابل پیگیری باشد.

مرجع‌ها

[۱] صنیعی آباده، م.، محمودی، س.، طاهرپرور، م. (۱۳۹۱). داده-کاوی کاربردی، نیاز دانش، شابک: ۸-۱۲-۶۴۸۱-۶۰۰-۹۷۸.

[2] Rachel, B., (2001). “Introductory Guide to Crime Analysis and Mapping”, Report to the Office of Community Oriented Policing Services, Cooperative Agreement #97-CK-WXK-004.

[3] Adderley, R., Musgrove, P.B., (2001). “General review of Police crime recording and investigation systems”. A user’s view. Policing: An International Journal of Police Strategies and Management, 24(1).

[4] Karlis, D., Meligkotsidou, L. (2007). “Finite Mixtures of Multivariate Poisson Distributions With Application”, Journal of Statistical Planning and Inference, No. 137, PP. 1942 – 1960.

[5] Buczak, L., Gifford, M., (2010). “Fuzzy Association Rule Mining for Community Crime Pattern Discovery”, ISI-KDD, Washington, D.C., USA, ACM, ISBN: 978-1-4503-0223-4/10/07.

[6] Moon, B., McCluskey, J.B., McCluskey, C.P., (2010). “General Theory of Crime and Computer Crime: An Empirical Test”, Journal Criminal Justice, No of. 38, PP. 767–772.

[7] Ozkan, K., (2004). “Managing Data Mining at Digital Crime Investigation”, Forensic Science International, No. 146, PP. S37–S38.

[8] Murtagh, F., Ganz, A., McKie, S. (2009), “The Structure of Narrative: The Case of Film Scripts, Pattern Recognition”, No. 42, PP. 302 – 312.

[9] Mande, U., Srinivas. Y., J.V.R.Murthy, J.V.R., “An Intelligent Analysis Of Crime Data Using Data Mining & Auto Correlation Models”, International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue 4, July-August 2012, pp.149-153, ISSN: 2248-9622, www.ijera.com.

[10] Corapcıoglu, A., Erdogan, S., (2004). “A Cross-Sectional Study on Expression of Anger and Factors Associated With Criminal Recidivism in Prisoners With Prior Offences”, Forensic Science International, No. 140, PP.167–174.

[11] Liu, H., Brown, Donald E., (2003). “Criminal Incident Prediction Using a Point-Pattern-Based Density Model”, International Journal of Forecasting, No. 19, PP. 603–622.

[12] D'Alessio, S.J., Stolzenberg, L., (2010). “Do Cities Influence Co-Offending?”, Journal of Criminal Justice, No. 38, PP. 711–719.

[13] Deadman, D. (2003), “Forecasting Residential Burglary”, International Journal of Forecasting, No. 19, PP. 567–578.

[14] Freilich, J.D., Pridemore, W.A. (2007). “Politics, Culture, and Political Crime: Covariates of Abortion Clinic Attacks in the United States”, Journal of Criminal Justice, No. 35, PP. 323–336.

[15] Xue, Y. & Brown , D.E. (2006), Spatial Analysis with Preference Specification of Latent Decision Makers for Criminal Event Prediction, Decision Support Systems, No. 41, PP. 560– 573.

[16] Hadjidj, R., Debbabi, M., Lounis, H., Iqbal, F., Szporer, A., Benredjem, D., (2009). “Towards an Integrated E-Mail Forensic Analysis Framework”, digital investigation, No. 5, PP. 1 2 4 – 1 3 7.

[17] Malathi, A., Santhosh, B., (2011). “An Enhanced Algorithm to Predict a Future Crime using Data Mining”, International Journal of Computer Applications (0975 – 8887), Volume 21.

[18] Li, SH.T., Kuo, SH.CH., Tsai, F.CH., (2010), “An Intelligent Decision-Support Model Using FSOM and Rule Extraction for Crime Prevention”, Expert Systems with Applications, No. 37, PP. 7108–7119.

[19] Oatley, G.C., Ewart, B.W. (2003), “Crimes Analysis Software: ‘Pins in Maps’, Clustering and Bayes Net Prediction”, Expert Systems with Applications, No. 25, PP. 569–588.

[20] D'Alessio, S.J., Stolzenberg, L., (2010). “Do Cities Influence Co-Offending?”, Journal of Criminal Justice, No. 38, PP. 711–719.

[21] www.gitashenasi.com

[22] http://data.police.uk/