طراحی الگوی داده‌کاوی پیشنهادی به‌منظور شناسایی مجرمان

تاریخ دریافت: ۱۳۹۵/۰۴/۲۳ تاریخ پذیرش: ۱۳۹۵/۰۶/۲۱

امیر مانیان،[۱] محمد جمالو،[۲] معصومه بیدل[۳]

از صفحه ۱۰۹ تا ۱۲۸

چکیده

زمینه و هدف: این پژوهش بر آن است تا با بهره‌گیری از الگوریتم‌های داده‌کاوی به تحلیل داده‌های ثبت‌شده در بانک اطلاعاتی پلیس مربوط به‌دستگیرشدگان توسط گشت‌های انتظامی تهران بزرگ در سه‌ماهه اول سال ۱۳۸۹ بپردازد و با استفاده از آنها، الگویی طراحی شود که به شناسایی مجرمان واقعی از بین انبوه متهمان دستگیرشده اقدام کند. این الگو می‌تواند به‌عنوان یک سامانه تصمیم‌یار در اختیار کارشناسان انتظامی قرار گیرد تا فرآیند شناسایی و دستگیری مجرمان واقعی با سرعت و دقت بیشتری انجام شود.

روش‌شناسی: این پژوهش از نوع پژوهش‌های داده‌محور بوده و بر اساس یک فرایند استاندارد داده‌کاوی CRISP-DM، داده‌های دستگیرشدگان که شامل متغیرهای جمعیت‌شناختی متهمان و کلانتری محل دستگیری است، پس از یکپارچه‌سازی و پالایش، با استفاده از الگوریتم‌های CHAID,CRT C5.0 و شبکه عصبی MLP مدل‌سازی شدند.

یافته‌ها: الگوریتم C5.0 در فن درخت تصمیم نتایج بهتری را به لحاظ دقت شناسایی مجرمان واقعی نسبت به سایر الگوریتم‌های درخت تصمیم، مانند CHAID, CRT دارد؛ اما نسبت به الگوی طراحی‌شده توسط شبکه عصبی MLP دقت کمتری دارد.

نتایج: با استفاده از الگوریتم‌های درخت تصمیم، در مجموع ۱۹ قانون کشف و ارائه شد. برای بررسی این قوانین، نشست خبرگان تشکیل شد و در نهایت از ۱۹ قانون استخراج‌شده، ۳ قانون مرتبط با موضوع مورد پژوهش شناخته شده و مورد تأیید قرار گرفت.

واژه‌های کلیدی: داده‌کاوی، شبکه عصبی، درخت تصمیم، جرم، بانک اطلاعاتی.

۱- دانشیار دانشکده مدیریت دانشگاه تهران، ایران، amanian@ut. ac. ir

۲- دانشجوی دکتری مدیریت IT، دانشگاه آزاد اسلامی، واحد علوم وتحقیقات تهران، ایران (نویسنده مسئول)، jamaloo2006@gmail. com

۳- دانشجوی دکتری مدیریت IT، دانشگاه آزاد اسلامی، واحد علوم وتحقیقات تهران، ایران، shbidel@yahoo.com

مقدمه

امروزه سازمان‌ها تقریباً از همه فرایندهای خود، داده جمع‌آوری می‌کنند و نیاز دارند تا از داده‌های عملیاتی خود درس بگیرند؛ اما کمتر از آنها استفاده می‌کنند، به‌عبارت دیگر سازمان‌ها در حال غرق شدن در داده‌ها هستند؛ اما همچنان تشنه دانش هستند. بانک‌های اطلاعاتی عظیم، زمانی ارزش سازمانی خواهند یافت که به ورودی‌های نظام مدیریت دانش سازمان تبدیل شده و با استفاده از الگوریتم‌های داده‌کاوی، نسبت به کشف قوانین و الگوها در میان انبوه داده‌ها پرداخته شود و در راستای تحقق اهداف پلیس مورد بهره‌برداری قرار گیرند (چو،[1] ۱۳۸۳: ۲۰۴). بنابراین داده‌های جمع‌آوری‌شده را می‌بایست طبقه‌بندی و تبدیل به دانش کرد تا سازمان‌ها بتوانند متناسب با آن عمل کنند. در اینجا است که ضرورت استفاده از داده‌کاوی[2] قابل طرح است تا بتوان با استفاده از آن به استخراج دانش موجود در انبوه داده‌های ذخیره‌شده (مانند قوانین و مقررات، الگوها، محدودیت‌ها) پرداخت (شهرابی و شکورنیاز، ۱۳۸۸: ۲۸۰).

سازمان پلیس نیز به‌عنوان یک سازمان پیشرو در عرصه فناوری اطلاعات، با داشتن سامانه‌های اطلاعاتی رایانه‌ای برخط،[3] با حجم زیادی از داده‌های اطلاعات عملیاتی در تمامی پلیس‌های تخصصی مواجه است. می‌توان با داده‌کاوی بر روی این داده‌های ذخیره‌شده و بهره‌گیری از نظر خبرگان، به دانش بسیار باارزش در حوزه‌های مختلف پلیس دست یافت. با توجه به اینکه یکی از مهم‌ترین وظایف سازمان پلیس، پیشگیری از جرم بوده و به‌طور مستمر داده‌های عملیاتی بسیار باارزشی از ویژگی‌های مجرمان و شرح جرم‌های صورت گرفته در حال ذخیره شدن در بانک‌های اطلاعاتی پلیس است، استفاده از الگوریتم‌های داده‌کاوی برای ایجاد الگوهای تحلیلی و پیشگیرانه، امری ضروری به نظر می‌رسد.

علی‌رغم اهمیت این موضوع و همچنین وجود داده‌های بسیار باارزش و باکیفیت در بانک اطلاعاتی پلیس، تاکنون هیچ‌گونه مدل‌سازی عملیاتی، برای پیشگیری از جرم انجام نشده است و مقالات معدودی به ارائه تعاریف نظری از مسئله پرداخته‌اند.

1- Cho
2- Data Mining
3- Online

این پژوهش بر آن است تا با بهره‌گیری از الگوریتم‌های داده‌کاوی به تحلیل داده‌های ثبت‌شده در بانک اطلاعاتی پلیس مربوط به‌دستگیرشدگان توسط گشت‌های انتظامی تهران بزرگ در سه‌ماهه اول سال ۱۳۹۲ بپردازد و با استفاده آنها الگویی طراحی شود تا بتوان با بهره‌گیری از آن، به شناسایی مجرمان واقعی از بین انبوه متهمان دستگیرشده اقدام کرد.

بیان مسئله

همواره گشت‌های پلیس در سراسر دنیا در حال جستجوی مجرمان واقعی هستند و بر این اساس اقدام به بازداشت افراد مشکوک و متهم به ارتکاب جرم می‌کنند. با توجه به اینکه این دستگیری‌ها تنها بر اساس تجربیات مأمور و شواهد ظاهری انجام می‌شود، باعث صَرف هزینه و وقت زیاد برای سازمان پلیس در کشف مجرمان واقعی شده است. در اکثر کشورهای دنیا، تمامی اطلاعات مرتبط با این دستگیری و بازجویی‌ها در پایگاه‌های اطلاعاتی پلیس ثبت و ذخیره می‌شوند و با استفاده از الگوریتم‌های داده‌کاوی، الگوهایی برای کوچک‌تر کردن دایره بررسی‌های پلیسی و شناسایی مجرمان واقعی با صرف زمان و هزینه بسیار کمتر طراحی می‌شود. همچنین از این داده‌های گران‌قیمت برای طراحی الگوهای پیشگیرانه جرم استفاده می‌شود. با توجه به‌وجود داده‌های عملیاتی در بانک اطلاعاتی پلیس، این تحقیق بر آن است تا به طراحی الگوی شناسایی مجرمان واقعی و ارائه قوانین پیشگیرانه اقدام کند.

مبانی نظری پژوهش

برای تبیین نظری پیشگیری از جرم در مقاله حاضر، لازم است تا به بررسی مطالب از دو بعد جرم‌شناسی و بعد فنی الگوریتم‌های داده‌کاوی پرداخته شود.

الف) بعد جرم‌شناسی: در قانون مجازات اسلامی تعریفی از جرم ارائه نشده است، فقط در مادهٔ ۲ قانون مجازات اسلامی در بیان اوصاف جرم آمده است: «هر فعل یا ترک فعلی که در قانون برای آن مجازات تعیین شده باشد جرم تلقی می‌شود.»

بر اساس نظر علمای حقوق کیفری، جرم به سه دسته کلی تقسیم می‌شود:

۱- جرم کیفری: جرم کیفری به معنای عام، عبارت است از هر فعلی که به‌موجب قوانین کیفری انجام دادن و یا ترک آن با مجازات مقرّر توأم باشد؛ مانند قتل، کلاه‌برداری، سرقت و غیره (اردبیلی، ۱۳۸۶: ۲۲۶).

۲- جرم مدنی: به فعلی اطلاق می‌شود که من غیر حق، زیانی به دیگری وارد و فاعل را به جبران آن ملتزم کند و ممکن است نصّ خاصّی در قانون نداشته باشد (اردبیلی، ۱۳۸۶: ۱۲۱).

۳- **جرم انتظامی**: تخلف انتظامی عبارت است از نقض مقررات صنفی یا گروهی که اشخاص به تبع عضویت در گروه، آن را پذیرفته‌اند. در واقع، جامعهٔ کوچکی مانند کانون‌های صنفی وکلا، سردفتران، پزشکان و... مانند جامعهٔ بزرگ، متکی به اصول و مقرراتی است که حافظ نظم و بقای گروه یا اتحادیهٔ صنفی و حرفه‌ای است (اردبیلی، ۱۳۸۶: ۱۲۳).

ب) الگوریتم‌های داده‌کاوی: فرآیند به خدمت گرفتن یک روش‌شناسی رایانه‌ای که با استفاده از فن‌های مختلف مستقیماً از داده‌ها دانش استخراج می‌کند، داده‌کاوی نامیده می‌شود (کاتان،[1] ۱۳۷۶: ۴۶۷). داده‌کاوی عبارت است از فرایند استخراج اطلاعات معتبر، از پیش ناشناخته، قابل فهم و قابل اعتماد از پایگاه داده‌های بزرگ و استفاده از آن در تصمیم‌گیری فعالیت‌های تجاری مهم (جفری و سیفرت[2]، ۱۳۸۳: ۸۵).

شبکه‌های عصبی: شبکه‌های عصبی، روشی است که قصد دارد با استفاده از الگوهای ریاضی و توان رایانه، برخی از جنبه‌های سادهٔ مغز انسان را شبیه‌سازی کند. شبکه‌های عصبی به‌صورت یکی از بخش‌های پیچیده مغز انسان، به‌عنوان یک ساختار یادگیری غیرقابل درک، مشهور شده است. این ساختار پیچیده از مجموعه‌ای از نرون‌ها به‌وجود آمده است که خود نرون‌ها ساختار ساده‌ای داشته، ولی شبکه اتصال این نرون‌ها وظایف یادگیری بسیار پیچیده‌ای را به انجام می‌رساند (شهرابی و شکورنیاز، ۱۳۸۸: ۲۳۰).

درخت تصمیم: درخت تصمیم یکی از ابزارهای قوی و متداول برای دسته‌بندی و پیش‌بینی است. درخت تصمیم برخلاف شبکه‌های عصبی به تولید قاعده می‌پردازد.

1- Kattan
2- Jefry, Seifert

ساختار درخت تصمیم یک ساختار درختی، شبکه فلوچارت دارد. بالاترین گره در درخت، گره ریشه است و گره‌های برگ، دسته‌ها یا توزیع دسته‌ها را نشان می‌دهند (شهرابی و ذوالقدر شجاعی، ۱۳۸۸: ۲۴۸)؛ درحالی‌که در شبکه‌های عصبی تنها نتیجه پیش‌بینی بیان می‌شود و چگونگی به‌دست آمدن آنها در خود شبکه پنهان می‌ماند.

پیشینه پژوهش

ضمن بررسی‌های به‌عمل‌آمده از منابع اطلاعاتی موجود در رابطه با موضوع تحقیق، درمی‌یابیم که تاکنون هیچ‌گونه الگوسازی عملیاتی داخلی، برای پیشگیری از جرم انجام نشده است و مقالات معدودی به ارائه تعاریف نظری از مسئله پرداخته‌اند. در ادامه به معرفی تعدادی از تحقیقات انجام‌شده در سایر کشورها اقدام خواهد شد.

ازکان (۲۰۰۴: ۳۷) جرم‌شناسی را در دو حوزه اقدامات قبل از وقوع جرم[1] و بعد از وقوع جرم دسته‌بندی کرده است. منظور از اولی، پیش‌بینی و پیش‌گیری از ارتکاب جرم و منظور از دومی، بررسی و کشف شواهد جرم پس از ارتکاب آن است. بررسی مقالات منتشره در زمینه کاربرد فن‌های داده‌کاوی در این‌گونه مسائل نشان می‌دهد که این حوزه در مهروموم‌های اخیر، مورد توجه پژوهش‌گران قرار گرفته است. در این میان فن‌های پیش‌بینی، بیشترین حجم مقالات را به خود اختصاص داده‌اند (احمدوند و آخوندزاده، ۸:۱۳۸۹).

پژوهشی توسط کراپسیوگلو و همکاران (۲۰۰۴) در مورد خصوصیات جمعیت‌شناختی[2] و اخلاقیِ مجرمانی که دوباره مرتکب جرم شده‌اند، در شهر ازمیت[3] انجام شده است. هدف از انجام این پژوهش، کشف خصوصیاتی بود که منجر به ارتکاب مجدد جرم می‌شد. در این مطالعه تمامی زندانیان زندان ازمیت مورد بررسی قرار گرفتند. در این راستا تمام اطلاعات جمعیت‌شناختی و روان‌شناختی زندانیان، جمع‌آوری شد و از آنها خواسته شد که اطلاعات مربوط به وضعیت محکومیت، ماهیت جرم، تاریخچه ارتکاب جرم، استعمال دخانیات و الکل و وضعیت ارتباط آنها با سایر زندانیان و کارکنان زندان را از طریق پرسش‌نامه‌ای که به این منظور طراحی شده بود،

1- Precrime
2- Demographic
3- Izmit

در اختیار محققان قرار دهند. بر اساس نتایج به‌دست‌آمده، افرادی که زودتر نسبت به دیگران عصبانی و خشمگین می‌شوند، نسبت به سایر زندانیان بیشتر مرتکب جرم شده‌اند (کراپسیوگلو و همکاران، ۲۰۰۴: ۱۶۷).

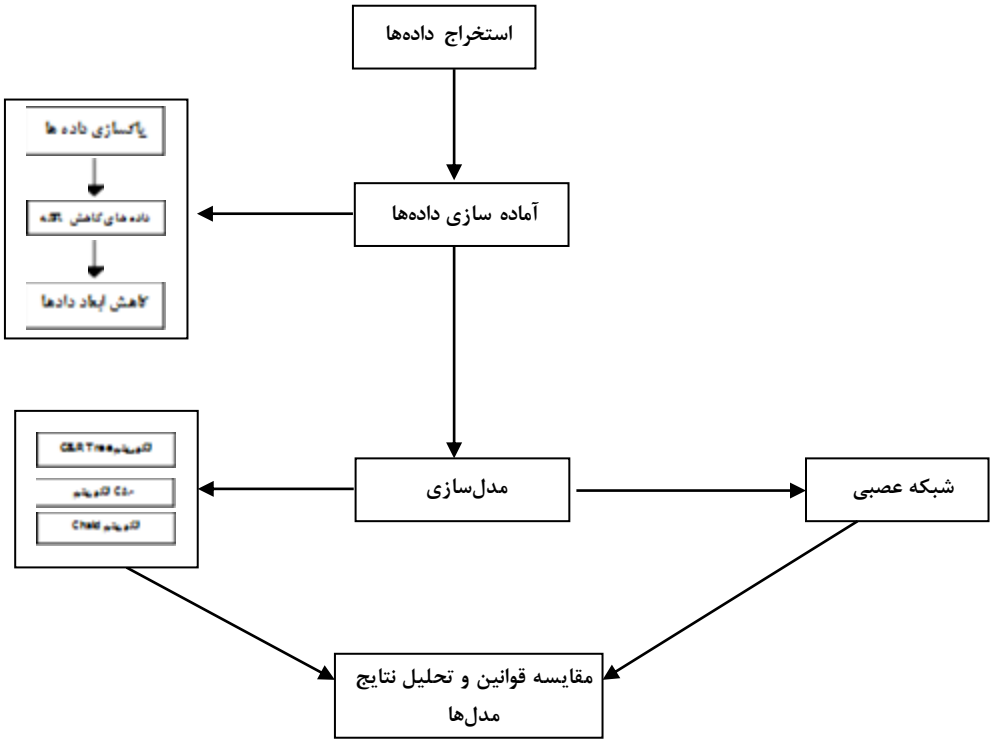
یکی از تحقیقات مشترک انجام‌شده بین پلیس انگلستان و گروه روان‌شناسی دانشگاه ساندرلند[1]، پژوهشی است که توسط اُتلی و همکاران (۲۰۰۳) انجام شده است. هدف اصلی از انجام این پروژه، کمک به نیروی پلیس در رسیدگی به جرایمی مانند دزدی از منازل بود که با نرخ زیادی صورت می‌گرفت. در این راستا نرم‌افزاری طراحی شد که بر اساس اطلاعات دقیق، متخلفان محل و زمان ارتکاب جرم به تدوین راهبردهای کوتاه‌مدت، میان‌مدت و بلندمدت برای کاهش جرایم کمک می‌کرد. در طراحی این نرم‌افزار از فن‌های داده‌کاوی اعم از تحلیل‌های آماری و روش‌های پیش‌بینی و هم‌چنین مباحث مرتبط با روان‌شناسی و جرم‌شناسی استفاده شده بود. از سری‌های زمانی برای پیش‌بینی تاریخ وقوع دزدی‌هایی که در یک روز تکرار می‌شد، استفاده شد. بر اساس نتایج به‌دست‌آمده، با این روش می‌توان وقوع دزدی مشابه در یک روز را پیش‌بینی کرد (اتلیو و همکاران، ۲۰۰۳: ۱۵۶).

مون و همکاران (۲۰۱۰) از رگرسیون برای پیش‌بینی جرایم رایانه‌ای استفاده کردند. بر اساس نتایج به‌دست‌آمده، میزان ساعات استفاده از رایانه و عضویت در گروه‌ها و شبکه‌های اینترنتی، میزان جرایم رایانه‌ای را افزایش داده و به‌عنوان متغیرهای اصلی پیش‌بینی‌کننده میزان جرایم معرفی شدند (مون و همکاران، ۲۰۱۰: ۷۶۷).

مدل مفهومی پژوهش

براساس مبانی نظری ارائه شده و همچنین پیشینه تحقیقات مرتبط با موضوع که پیش از این به آنها اشاره شد، مدل مفهومی تحقیق به شرح زیر ارائه می‌شود.

1- Sunderland

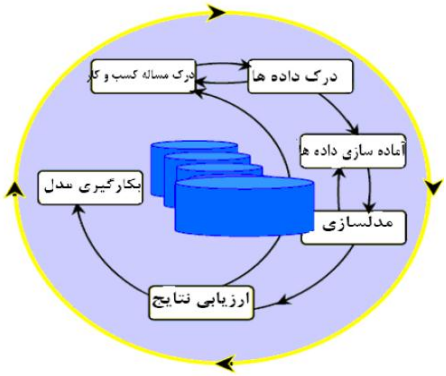


شکل شماره (۱): مدل مفهومی پژوهش

روش‌شناسی پژوهش

با توجه به ماهیت پژوهش که استفاده از داده‌کاوی برای شناسایی مجرمان واقعی از بین متهمان دستگیرشده در گشت‌های انتظامی تهران در سه‌ماهه اول سال ۱۳۹۲ است، این پژوهش از نوع داده‌محور بوده و پایه اصلی پژوهش حاضر بر کشف دانش از پایگاه داده‌های بانک مورد مطالعه نهاده شده است. نگرش این پژوهش از پایین به بالا است؛ به این شکل که از کار با داده‌ها شروع شده و سعی بر آن است تا مواردی را که قبلاً آگاهی نسبت به آنها وجود نداشته است، کشف کرده و برای آنها قوانینی ساخته شود؛ بنابراین از روش کشف دانش استفاده شده است.

بر این اساس، استاندارد جهانی CRISP-DM[1] در حوزه داده‌کاوی، برای انجام فرایند پژوهش و ساختار اجرایی در این پژوهش مورد استفاده قرار گرفته است. الگوی جدید استاندارد CRISP-DM مطابق شکل یک است که بر اساس آن، روش پژوهش تشریح می‌شود.



شکل شماره (۲): فرآیند مدل داده‌کاوی CRISP-DM

۱- **شناخت مسئله کسب‌وکار:** در این فاز از فرایند، ابتدا اهداف اصلی کسب‌وکار تعیین شد که اصلی‌ترین هدف کسب‌وکار (پلیس پیشگیری تهران بزرگ) در این مقاله، پیدا کردن روابط پنهان در بانک اطلاعات متهمان دستگیرشده توسط گشت‌های انتظامی تهران بزرگ در سه‌ماهه اول سال ۱۳۸۹ است که داده‌هایشان در سامانه رایانه‌ای پلیس ثبت شده است. این کار بر اساس کشف الگوهای پنهان موجود بین ویژگی‌های جمعیت‌شناختی، سابقه و محل دستگیری متهمان دستگیرشده است.

۲- **درک داده‌ها**[2]**:** جامعه آماری پژوهش متشکل از داده‌های مربوط به ۳۷۰۰۰ نفر متهم است که در طی سه‌ماهه اول سال ۱۳۸۹ توسط گشت‌های انتظامی تهران بزرگ دستگیرشده و اطلاعات آنها در بانک اطلاعاتی پلیس ثبت شده است. لازم به ذکر است برای رعایت محرمانگی اطلاعات، داده‌ها به‌صورت رمزنگاری‌شده

1- CRoss Industry Standard Process for Data Mining
2- Data understanding

دریافت شده و رکوردهای استفاده‌شده در پژوهش به‌صورت رندم از مجموع داده‌ها اخذ شد.

۳- **آماده‌سازی داده‌ها**[1]: مرحله آماده‌سازی داده‌ها یکی از مهم‌ترین مراحل در داده‌کاوی است که شامل اقداماتی مانند پاک‌سازی داده‌ها، کاهش ابعاد داده‌ها و تبدیل و یکپارچه کردن آنها است. اگرچه این مرحله به ظاهر ساده‌ترین مرحله کار به نظر می‌رسد؛ اما در واقع یکی از پیچیده‌ترین، حساس‌ترین و زمان‌برترین مراحل فرایند داده‌کاوی است.

مرحله آماده‌سازی داده‌ها در فرایند داده‌کاوی، مرحله زمان‌بر و بااهمیتی است که حدوداً ۶۰ تا ۷۰ درصد زمان انجام کل فرایند داده‌کاوی را به خود اختصاص می‌دهد (هو[2]، ۱۳۷۸). یکی از بزرگ‌ترین مشکلاتی که صحت و دقت داده‌های بانک‌های اطلاعاتی بزرگ را به چالش می‌کشد، وجود داده‌های متناقض[3]، مفقود[4] و مغشوش[5] است. صاحب‌نظران علم داده‌کاوی، یکی از علل عمده کیفیت پایین نتایج حاصل از داده‌کاوی را، کیفیت پایین داده‌های ورودی و عدم توجه به مراحل آماده‌سازی داده‌ها می‌دانند.

به علت حجم زیاد داده‌های دریافتی از متهمان و وجود اطلاعات در چندین جدول مختلف ابتدا بر اساس مسئله تعریف‌شده، داده‌های فاقد متغیر مورد نظر محقق از پایگاه داده حذف شد، سپس اطلاعات مورد نیاز برای هر متهم در تمامی جداول جمع‌آوری شده و در یک جدول ذخیره شد، به‌این‌ترتیب هر متهم در جدول نهایی دارای یک رکورد بوده که تمامی متغیرهای مورد نظر محقق برای آن موجود بود.

پس از انجام مرحله آماده‌سازی داده‌ها، یازده مشخصه به شرح جدول شماره یک، برای الگوسازی استفاده شد. از ۷۰ درصد جامعه آماری از آنها صرف‌نظر شد و نهایتاً فیلدهای ذکر شده در ادامه به‌عنوان متغیرهای توصیف‌کنندهٔ هر متهم مورد استفاده قرار گرفت، فیلدی با عنوان Target به مجموع فیلدها اضافه شد که دارای مقدار باینری (۱ و ۰) است که مقدار ۱ نشان‌دهنده افراد مجرم و ۰ نشان‌دهنده افراد بی‌گناه است.

1- Data preparation
2- Ho
3- Inconsistent
4- Missing
5- Noisy

جدول شماره (۱): مشخصه‌های نهایی برای مدل‌سازی

ردیف	عنوان
۱	محل دستگیری
۲	استان محل سکونت
۳	استان محل تولد
۴	زمان دستگیری
۵	گروه سنی
۶	جنسیت
۷	وضعیت تأهل
۸	رشته تحصیلی
۹	مقطع تحصیلی
۱۰	شغل
۱۱	Target

۴- **الگوسازی:** در مرحله الگوسازی، روش‌های مختلف ساخت الگو، انتخاب و به کارگرفته می‌شوند. بر این اساس در این پژوهش از دو روش درخت تصمیم (دسته‌بندی) و شبکه عصبی استفاده شده است که در دو مدل اولیه و ثانویه، مدل‌سازی به شرح ذیل انجام شده است.

۴-۱) ایجاد الگو اولیه: در طراحی الگوی اولیه از الگوریتم‌های,CRT C5.0 CHAID درخت تصمیم، به‌طور جداگانه استفاده شد و در پایان سه الگوی داده‌کاوی مجزا از هریک از الگوریتم‌ها استخراج شد. خروجی الگو در این الگوریتم‌ها به دو صورت زیر استخراج شد:

اول اینکه خروجی‌ها به‌عنوان یک طبقه‌بندی کننده داده‌ها مورد استفاده قرار گیرند و الگویی صحیح و دقیق بر اساس معیارها طراحی شد. دوم اینکه قوانینی برای پیش‌بینی آینده به‌دست آمد.

۴-۲) ایجاد مدل ثانویه: برای ارزیابی خروجی‌های الگوی ایجادشده توسط الگوریتم‌های درخت تصمیم، با استفاده از شبکه عصبی MLP، الگوی ثانویه ایجاد شد. در این مرحله نتایج به‌دست‌آمده از الگوی شبکه عصبی MLP با نتایج هریک از الگوریتم‌های درخت تصمیم به‌صورت مجزا مورد مقایسه قرار گرفت تا با استفاده از آن الگوریتمی که دارای بالاترین دقت است به‌عنوان الگوی خروجی انتخاب شود.

۴-۳) بررسی قوانین کشف‌شده در نشست خبرگان: برای ارزیابی قوانین به‌دست‌آمده از الگوها (در مدل اولیه) نتایج حاصل از آنها، به اطلاع خبرگان پلیس پیشگیری تهران بزرگ رسید. این افراد از بین قوانین استخراجی، قوانینی که مورد تأیید بودند را مشخص کردند. در انتها قوانین و الگوهای موجود در داده‌های متهمان بر اساس طبقات تعریف‌شده استخراج شده و برای بررسی در نشت خبرگان مطرح شد.

ارزیابی اعتبار الگوها

در نهایت به‌دلیل آنکه روش ارائه‌شده در هر پژوهشی باید به لحاظ اعتبار، مورد سنجش قرار گیرد، در این پژوهش نیز با عنایت به «داده محور» بودن، روش اعتبارسنجی به این صورت اجرا شد که داده‌ها به دو مجموعه داده‌های آموزشی و داده‌های آزمایشی تقسیم شدند. هدف از این کار این بود که با الگوسازی داده‌های آموزشی، الگوریتم انتخابی و دانشی حاصل می‌شود؛ ولی اینکه نتایج حاصله تا چه میزان دارای اعتبار هستند باید توسط نتایج داده‌های جدید و قدرت پیش‌بینی الگوریتم در مورد داده‌هایی که تاکنون با آن مواجه نبوده، آزمون شوند؛ از این‌جهت داده‌های آزمون به‌عنوان داده‌های ناظر به الگوریتم، داده شدند و نتایج حاصله میزان دقت و صحت الگو را ارزیابی کرد.

یافته‌های پژوهش

الف) یافته‌های توصیفی

با استفاده از الگوریتم‌های درخت تصمیم، در مجموع ۱۹ قانون کشف و ارائه شد. برای بررسی و تأیید این قوانین نشست خبرگان تشکیل شد و در نهایت از ۱۹ قانون استخراج‌شده، سه قانون مورد تأیید خبرگان قرار گرفت. این قوانین عبارت‌اند از:

۱- متهمانی که دارای مدارک تحصیلی دیپلم و زیر دیپلم، در سنین ۲۱ سال و کمتر از آن بوده و مجرد هستند و محل تولد آنها استان‌های آذربایجان غربی، چهارمحال و بختیاری، زنجان، مازندران بوده و در حومه شهر تهران سکونت داشتند بیشتر از سایر افراد دچار جرم می‌شوند. این آمار در همین دسته از متهمان، اما با

ویژگی متأهل بودن دارای کاهش چشمگیری است، این موضوع به‌خصوص در استان آذربایجان غربی بیشتر قابل رؤیت است.

۲- متهمانی که دارای مدرک تحصیلی فوق‌دیپلم هستند، کمتر از سایر افراد دچار ارتکاب جرم می‌شوند.

۳- متهمانی که دارای مدرک تحصیلی لیسانس به بالا باشند، کمتر از سایر افراد مرتکب جرم می‌شوند.

تحلیل صورت گرفته توسط خبرگان این بود که چون افراد دیپلم و زیر دیپلم که دارای سن ۲۱ و کمتر هستند عموماً به کارهای سطح پایین جامعه مانند کارگری اشتغال دارند، فشارهای اجتماعی و نیازهای مالی، باعث می‌شود، بیشتر مستعد ارتکاب جرم باشند، ضمن اینکه متأهل بودن به‌دلیل به‌وجود آوردن تعهد در افراد، عامل بسیار مهم در کاهش ارتکاب به جرم در این دسته است.

ب) یافته‌های استنباطی

در این پژوهش، داده‌ها در دو کلاس طبقه‌بندی شده‌اند؛ کلاس متهمان واقعی و کلاس متهمان بی‌گناه. همان‌طور که در شکل شماره سه، مشاهده می‌کنید، نتیجه الگوسازی، در چهار حالت قرار می‌گیرد:

N1: دسته‌بندی صحیح متهمان واقعی؛

N2: دسته‌بندی صحیح متهمان بی‌گناه؛

M1: دسته‌بندی اشتباه متهمان کلاس واقعی در کلاس بی‌گناه؛

M2: دسته‌بندی اشتباه متهمان کلاس بی‌گناه در کلاس واقعی.

فرآیند داده کاوی[1]

مدل داده کاوی[2] CRISP-DM-ارزیابی دسته بندی

جدول مقایسه ای		حالت واقعی	
		Class 1	Class 2
حالت واقعی	Class 1	N1	M1
	Class 2	M2	N2

جدول مقایسه ای		حالت واقعی	
		Class 1 مثبت	Class 2 منفی
حالت واقعی	Class 1 مثبت	True مثبت	False منفی
	Class 2 منفی	False مثبت	True منفی

شکل شماره (۳): دسته‌بندی در استاندارد CRISP-DM

فرآیند داده کاوی

مدل داده کاوی CRISP-DM-ارزیابی دسته بندی

جدول مقایسه ای		نتیجه مدل	
		Class 1	Class 2
حالت واقعی	Class 1	N1	M1
	Class 2	M2	N2

- N1 دسته‌بندی صحیح متهمان واقعی
- N2 دسته‌بندی صحیح متهمان بی‌گناه
- M1 دسته‌بندی اشتباه متهمان کلاس واقعی در کلاس بی‌گناه
- M2 دسته‌بندی اشتباه متهمان کلاس بی‌گناه در کلاس واقعی

شکل شماره (۴): نتیجه الگوسازی در الگوهای باینری با توابع الگوریتم درخت تصمیم

رایج‌ترین معیارهای عملکردی که برای ارزیابی الگوریتم‌ها به کار می‌روند، دقت

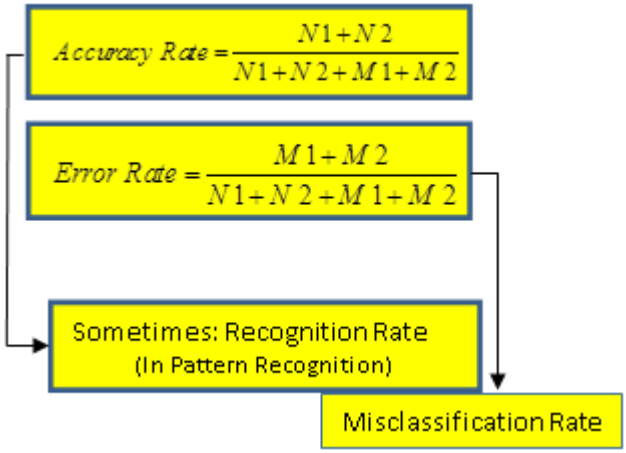
1- Data mining Process
2- Modeling choosing the evaluation function Classification

دسته‌بندی (نرخ دسته‌بندی صحیح) و نرخ خطا در تشخیص (نرخ دسته‌بندی اشتباه) است. نحوه محاسبه این نرخ‌ها در شکل شماره پنج، قابل مشاهده است.

فرآیند داده کاوی

مدل داده کاوی CRISP-DM-ارزیابی دسته بندی

جدول مقایسه ای		نتیجه مدل	
		Clas 1	Class 2
حالت واقعی	Class 1	N1	M1
	Class 2	M2	N2



شکل شماره (۵): تعریف معیارهای ارزیابی توابع الگوریتم درخت تصمیم

با توجه به دسته‌بندی، احتمال وقوع دو اشتباه M1 و M2 وجود دارد که در حالت عادی ارتکاب هر دو اشتباه دارای ارزش منفی یکسانی است؛ زیرا اگر M1>M2 باشد، سازمان پلیس به‌اشتباه یک مجرم دستگیرشده را آزاد می‌کند و درصورتی‌که M1<M2 باشد، سازمان پلیس به‌اشتباه یک بی‌گناه را دستگیر کرده است که هر کدام از آنها مسائل مرتبط به خود را در پی دارد. لیکن با توجه به اهداف پژوهش که شناسایی مجرمان واقعی و پیشگیری از جرم است، وقوع اشتباه M1 باعث تضعیف الگو و عدم تحقق اهداف پژوهش می‌شود؛ بنابراین در این پژوهش شاخص جدیدی با عنوان Min|E2-E1| تعریف شده است. به‌این‌ترتیب الگوهایی که این شاخص در آنها حداقل باشد، دارای ارزیابی بهتری هستند. در این پژوهش برای مقایسه نتایج دسته‌بندی

الگوریتم‌های درخت تصمیم و شبکه عصبی، علاوه بر شاخص نرخ دسته‌بندی صحیح و نرخ تشخیص اشتباه که توضیح داده شد از شاخص‌های حساسیت، نسبت تعیین و دقت که به‌صورت روابط زیر تعریف شده‌اند، نیز برای مقایسه استفاده شده است.

مدل ارزیابی ضرر نرخ صحت[1]

جدول مقایسه ای		نتیجه مدل	
		مثبت	منفی
حالت واقعی	مثبت	TP	FN
	منفی	FP	TN

سایر شروط مهم و مفید برای ارزیابی:

حساسیت[2] =recall= hit rate= TP / (TP + FN)

نرخ برتری ویژگی[3] = TN / (TN + FP)

دقت[4] = TP / (TP + FP)

شکل شماره (۶): نتایج دسته‌بندی

$$\text{حساسیت} = \frac{\text{تعداد متهمان واقعی صحیح طبقه‌بندی‌شده}}{\text{کل متهمان واقعی دستگیرشده}}$$

$$\text{نسبت تعیین} = \frac{\text{تعداد متهمان بی‌گناه صحیح طبقه‌بندی شده}}{\text{کل متهمان بی‌گناه دستگیرشده}}$$

$$\text{دقت} = \frac{\text{تعداد دسته‌بندی صحیح متهمان واقعی}}{\text{تعداد کل متهمان دستگیرشده}}$$

الف) مقایسه نتایج الگوریتم‌های درخت تصمیم و شبکه عصبی

1- Model Evaluation, Disadvantage of Accuracy Rate
2- Sensitivity
3- Specificity
4- Precision

جدول شماره (۲): مقایسه نتایج شاخص‌ها

شاخص	C5.0	C&R Tree	CHAID	شبکه عصبی
دقت کلی در دسته‌بندی	٪۸۱/۹۴	٪۷۹/۴۳	٪۷۹/۵	٪۸۰/۳۴
دقت در دسته‌بندی متهمان بی‌گناه	٪۷۲/۹۵	٪۵۶/۷	٪۶۲/۲۳	٪۷۲/۸۸
دقت در دسته‌بندی متهمان واقعی	٪۸۷/۱۷	٪۹۲/۶۶	٪۸۹/۵۵	٪۸۴/۶۸
نرخ خطا در تشخیص متهمان بی‌گناه	٪۲۷/۵	٪۴۳/۲۹	٪۳۷/۷۷	٪۲۷/۱۲
نرخ خطا در تشخیص متهمان واقعی	٪۱۲/۸۲	٪۷/۴۴	٪۱۰/۴۵	٪۱۵/۳۲

نتایج الگوهای الگوریتم درخت تصمیم به‌طور خلاصه در جدول شماره دو، آمده است. اکنون می‌خواهیم نتایج به‌دست‌آمده را با اهداف پژوهش، مقایسه کنیم. در برآورده کردن مهم‌ترین هدف که عبارت بود از بیشینه کردن دقت دسته‌بندی برای متهمان واقعی، الگوریتم C&R Tree از سایر الگوریتم‌های درخت تصمیم و شبکه عصبی بهتر عمل می‌کند.

الگوریتم C&R Tree، ۶۵/۷۳ درصد متهمان واقعی را تشخیص می‌دهد؛ درحالی‌که الگوریتم CHAID ٪۸۹/۵۵، الگوریتم C5.0 ٪۸۷/۱۷ و در شبکه عصبی ٪۸۴/۶۸، این دسته از متهمان را تشخیص می‌دهد. از نظر دومین معیار ارزیابی که عبارت بود از کمینه کردن نرخ خطا برای متهمان بی‌گناه، الگوریتم C5.0 با تشخیص ٪۷۲/۹۵ از متهمان بی‌گناه، برتری چشمگیری نسبت به سایر الگوریتم‌های درخت تصمیم دارد؛ اما شبکه عصبی با تشخیص ٪۷۲/۸۸، تفاوت بسیار ناچیزی با الگوریتم C5.0 دارد.

الگوریتم C5.0، ٪۲۷/۵ از متهمان بی‌گناه را در دسته متهمان واقعی قرار می‌دهد؛ درصورتی‌که الگوریتم CHAID، ٪۳۷/۷۷، الگوریتم C&R Tree، ٪۴۳/۲۹ از متهمان بی‌گناه و شبکه عصبی ٪۲۷/۱۲ را در دسته متهمان واقعی قرار می‌دهند که در این معیار نیز تشخیص دو الگوریتم C5.0 و شبکه عصبی، بسیار به هم نزدیک هستند؛ اما الگوریتم C5.0 برتری خوبی نسبت به سایر الگوریتم‌های درخت تصمیم دارد.

همان‌طور که بیان شد، علاوه بر شاخص‌های فوق‌الذکر، از شاخص‌های حساسیت، نسبت تعیین و دقت نیز استفاده می‌شود که در جدول شماره سه، قابل مشاهده است، این شاخص‌ها در تمامی الگوریتم‌ها، برای هر دو داده آموزشی و آزمون محاسبه شده‌اند.

جدول شماره (۳): مقایسه الگوریتم‌های درخت تصمیم

شاخص	نوع داده	C5.0	C&R Tree	CHAID	شبکه عصبی
حساسیت	آموزشی	٪۸۹/۳۱	٪۹۲/۳۳	٪۸۹/۳۷	٪۸۳/۸۸
	آزمون	٪۸۷/۱۷	٪۹۲/۶۶	٪۸۹/۵۵	٪۸۴/۶۸
نسبت تعیین	آموزشی	٪۷۸/۰۲	٪۵۸/۸	٪۳۶/۲۴	٪۷۳/۲۱
	آزمون	٪۷۲/۹۵	٪۵۷/۶۷	٪۶۲/۲۲	٪۷۲/۸۸
دقت	آموزشی	٪۵۶/۳۴	٪۵۸/۲۵	٪۵۶/۳۸	٪۵۲/۹۲
	آزمون	٪۵۵/۱	٪۵۸/۵۸	٪۵۶/۶۱	٪۵۳/۵۳

با جمع‌بندی نتایج شاخصه‌ها، می‌توان نتیجه گرفت که الگوریتم C5.0، برای شناسایی متهمان واقعی بر الگوریتم‌های C&R Tree و CHAID برتری دارد. این برتری در برخی از شاخص‌ها بسیار فاحش است و در برخی از شاخص‌ها چندان فاحش نیست؛ به‌طور مثال همان‌طور که بیان شد، شبکه عصبی در شاخص|M1-M2| برابر ۱۲ درصد و در شاخص دقت کلی دسته‌بندی، برابر ۸۰/۳۴ درصد بود.

نتایج این شاخص‌ها در الگوریتم C5.0، به ترتیب برابر ۱۸ درصد و ۸۱/۹۴ درصد است. با توجه به نتایج شاخص‌های فوق‌الذکر، به‌ویژه دو شاخص، |M1-M2| و دقت کلی دسته‌بندی که مهم‌ترین شاخص‌های این پژوهش هستند، نتایج شبکه عصبی و الگوریتم C5.0 همدیگر را تأیید می‌کنند و برتری خوبی نسبت به سایر الگوریتم‌ها دارند.

نتیجه‌گیری

آشکار است که موضوع داده‌کاوی جرم بسیار وسیع‌تر از آن است که بتوان به‌صورت جامع در قالب یک مقاله به آن پرداخت، لیکن تمام تلاش‌ها انجام شد تا با استفاده از داده‌های موجود و دو روش درخت تصمیم (دسته‌بندی) و شبکه عصبی، الگویی محدود برای پیش‌بینی جرم ارائه شود تا مورد استفاده مدیران سازمان پلیس و دستگاه قضایی قرار بگیرد. اگرچه برای استفاده کامل و اجرایی از این الگوها، لازم است تا داده‌های مناسب با مشخصه‌های بیشتر در اختیار محقق قرار داده شود و الگو توسط یک گروه فنی و نظارتی طراحی و ارزیابی شود.

اصولاً برخی از قوانین کشف‌شده بسیار واضح به نظر می‌رسد؛ اما یکی از ویژگی‌های داده‌کاوی کشف قوانینی است که در عین منطقی و واضح بودن، مورد توجه نبوده است

و کشف و ارائه این قوانین باعث می‌شود تا سازمان‌ها توجه بیشتری به حوزه‌های مغفول‌مانده کنند. به‌عنوان‌مثال یکی از قوانین کشف‌شده بیان می‌دارد که متهمانی که دارای مدرک تحصیلی فوق‌دیپلم باشند، کمتر از سایر افراد دچار ارتکاب جرم می‌شوند. با بررسی نظر خبرگان جرم‌شناسی، این قانون بسیار دقیق استخراج شده است و بسیار واضح است؛ زیرا اکثر افراد دارای مدرک فوق‌دیپلم، تکنسین‌هایی هستند که سریعاً جذب بازار کار شده و به‌دلیل داشتن شغل، درآمد و تعهد شغلی و اجتماعی، کمتر از سایر افراد جامعه مرتکب جرم و بزه در اجتماع می‌شوند.

پیشنهادها

الگوی ارائه‌شده در این پژوهش، الگوی جدیدی است و با توجه به اهمیت نقش پلیس در پیشگیری از جرم، لزوم توسعه و بهبود آن، ضروری به نظر می‌رسد. به‌عبارت‌دیگر این پژوهش شواهد کاربردی از تحلیل مسائل حوزه پیشگیری از جرم با استفاده از فن‌های داده‌کاوی را ارائه کرده است. با توجه به اینکه تاکنون تحقیقات داخلی جدی داده‌کاوی در حوزه کشف جرم در پلیس صورت نگرفته است، این پژوهش می‌تواند بسیار باارزش و مفید بوده و شروعی برای تحلیل‌های کاربردی‌تر در همین حوزه و سایر مسائل کشف جرم با استفاده از داده‌کاوی باشد. براین اساس پیشنهادهای زیر برای اجرا در تحقیقات آینده ارائه می‌شود:

- تعریف دسته‌های متنوع‌تر و استفاده از رویکرد فازی؛
- ارائه سازوکاری برای دخالت هزینه‌های مختلف دسته‌بندی در الگوریتم‌ها؛
- طراحی انبار داده مناسب و کدبندی‌شده در سازمان پلیس؛
- استفاده از داده‌های مناسب با مشخصه‌های بیشتر؛
- بهره‌گیری از یک گروه فنی و نظارتی برای طراحی و ارزیابی مدل.

منابع

- اردبیلی، محمدعلی (۱۳۸۶). حقوق جزای عمومی. تهران: نشر میزان.
- شهرابی، جمال؛ شکورنیاز، ونوس (۱۳۸۷). مفاهیم داده‌کاوی در اراکل 11g. تهران: انتشارات متالون.
- شهرابی، جمال؛ شکورنیاز، ونوس (۱۳۸۸). داده‌کاوی در SQL Server. تهران: جهاد دانشگاهی واحد صنعتی امیرکبیر.
- شهرابی، جمال؛ ذوالقدر شجاعی، علی (۱۳۸۸). داده‌کاوی پیشرفته مفاهیم و الگوریتم‌ها. تهران: جهاد دانشگاهی واحد صنعتی امیرکبیر.
- چو، جیم (۱۳۸۳). فناوری اطلاعات نیروی انتظامی. ترجمه فروهر دزفولیان، تهران: سازمان تحقیقات و مطالعات ناجا.
- جعفری، ادریس (۱۳۸۸). کاربرد داده‌کاوی در بررسی رفتار رانندگان متخلف در کلان‌شهرها. پایان‌نامه کارشناسی ارشد، دانشگاه تربیت مدرس.
- توکلی، احمد و همکاران (۱۳۸۹). به‌کارگیری فرایند داده‌کاوی برای پیش‌بینی الگوهای رویگردانی مشتری در بیمه. فصل‌نامه چشم‌انداز مدیریت بازرگانی، دانشگاه فردوسی مشهد، شماره ۴.
- احمدوند، علی‌محمد و همکاران (۱۳۸۹). چهارچوب کاربردی فن‌های داده‌کاوی در مدل‌سازی جرایم. دوماهنامه توسعه منابع انسانی پلیس، تهران، شماره ۳۰.
- Barson, P., Field, S., Davey, N., McAskie, G., Frank, R., (1996), the detection of fraud in mobile phone, networks. Neural Network World 6 (4), pp. 477–484.
- Breiman, L, /Friedenman, J. H. /Olshen, R. A. /Stone, C. J., (1984), Classification and regression trees. Monterey, CA: Wadsworth.), pp. 77–99.
- David J. HAND, Data Mining: Statistics and More? , December 2002.), pp. 22–48.
- Hand, D. J. /Henley, W. E., (1997). Statistical Classification Methods in Consumer Credit Scoring: a Review. In: J. R. Statist. Soc. A, 160, Part 3, pp. 523-542.
- Hoath, P., (1998). Telecoms fraud: The gory details. Computer Fraud & Security (January), pp. 223.
- Kattan, MW. /Cooper, RB. , (1997). The predictive accuracy of computer based classification decision techniques, a review and research directions. In: Omega, the International Journal of Management, Science, Vol. 26, No. 4, pp. 467-482.

- Liu, Y., (2001). New Issues in Credit Scoring Application. Research paper, Institute of Information Systems, University of Goettingen, Nr. 16, Göttingen.), pp. 477–484.
- Nong Ye (2003). The Hand book of Data Mining. New Jersey, LAWRENCE ERLBAUM ASSOCIATES.), pp. ۱۲۳–۱۳۰.
- Jeffery W. Seifert, Analyst in information science and Technology Policy, Data Mining: An Overview December 2004), pp. 180–284.
- U. M. Fayyad, G. Piatetsky-Shapiro, R. Uthurusamy (2003). Summary from the KDD-03 Panel -- Data Mining: The Next 10 Years, SIGKDD Explorations. Volume 5, Issue 2 – pp. 191-196.
- Ozkan, K. (2004). Managing Data Mining at Digital Crime Investigation, Forensic Science International, No. 146, PP. S37–S38.
- 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.
- 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.
- Moon, B., McCluskey, J. B. & McCluskey, C. P. (2010). General Theory of Crime and Computer Crime: An Empirical Test, Journal of Criminal Justice, No. 38, PP. 767–772.
- 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.