فصلنامه مطالعات مدیریت فناوری اطلاعات سال سوم، شماره ۱۲، تابستان ۹٤
صفحات ۱ تا ۲۰

رویکردی نوین به منظور کشف و تجزیه‌وتحلیل دانش پدیده‌های استثنایی با استفاده از داده‌کاوی

مسعود عابسی*
الهه حاجی گل یزدی**
حسن حسینی نسب***
محمدباقر فخرزاد****

چکیده

منطق یادگیری از استثنائات چالشی قابل‌توجه در حوزه داده‌کاوی است. استثنائات پدیده‌های نادری هستند که رفتاری مثبت و متفاوت از الگوهای اصلی و مورد انتظار موجود در پایگاه داده از خود بروز می‌دهند. ایجاد چارچوبی کارا برای افزایش اطمینان به پدیده‌های استثنایی در کشف دانش و یادگیری مؤثر از آن حائز اهمیت است. در این پژوهش، الگویی بر اساس تئوری استثنائات و تئوری اطلاعات ارائه شده است تا چالش‌های پیش روی داده‌کاوی داده‌های استثنایی را برطرف نماید. نخست از تابع آنتروپی رنی برای شناسایی استثنائات استفاده و سپس با به‌کارگیری رویکرد یادگیری پایین به بالا، بر مبنای الگوریتم پیشنهادی RISE ارتقا یافته، قوانین حاکم بر بروز رفتار استثنایی استخراج می‌گردد. به‌منظور تعیین کارایی مدل پیشنهادی، کشف سهام استثنایی و یادگیری رفتار آن‌ها مورد بررسی قرار گرفته است. از مجموع ۱۳۳٤ سهم مورد بررسی ۳٦ سهم رفتار استثنایی داشته‌اند که رفتار آن‌ها در قالب سه قانون مشخص شده

*استادیار گروه مهندسی صنایع، دانشکده فنی مهندسی، دانشگاه یزد، یزد، ایران
** دانشجوی دکتری مهندسی صنایع، دانشکده فنی مهندسی، دانشگاه یزد، یزد، ایران (نویسنده مسئول) elahehajigol@gmail.com
*** دانشیار گروه مهندسی صنایع، دانشکده فنی مهندسی، دانشگاه یزد، یزد، ایران
**** استادیار گروه مهندسی صنایع، دانشکده فنی مهندسی، دانشگاه یزد، یزد، ایران

تاریخ دریافت: ۹٤/۲/۱٤ **کد مقاله:۱۰۲٦-۱۵۰۵** **تاریخ پذیرش: ۹٤/۹/۲۲**

است. ارجحیت نتایج حاصل از مدل پیشنهادی نسبت به نتایج به‌دست‌آمده از به‌کارگیری الگوریتم‌های معمول یادگیری بیانگر کارایی مدل ارائه‌شده است.

کلیدواژگان: داده‌کاوی، تئوری استثنائات، تئوری اطلاعات، الگوریتم یادگیری پایین به بالا، پدیده‌های استثنایی

مقدمه

داده‌کاوی فرایند کشف و یادگیری الگوها و ارتباطات موجود در انباره داده است. منطق یادگیری رفتار استثنائات به‌عنوان یک چالش در حیطه داده‌کاوی، زمانی به عرصه ظهور رسید که یادگیری ماشین از یک دانش ابتدایی به یک فنّاوری کاربردی ارتقا یافت (چاولا و همکاران[1]، ۲۰۰٤). داده‌های موجود در یک پایگاه داده از لحاظ ماهیت به سه دسته داده نرمال، استثنایی و پرت تقسیم می‌شوند. داده‌های نرمال از الگوهای اصلی و عمومی موجود در یک پایگاه داده تبعیت می‌نمایند، درحالی‌که داده‌های استثنایی و پرت رفتاری متفاوت از الگوهای عمومی مورد انتظار موجود در پایگاه داده دارند. داده‌های استثنایی به دلیل رفتار مافوق تصور خوب یا تغییر مثبت در رفتار سیستم حاصل می‌شود. منشأ بروز داده‌های استثنایی وجود خطا در سیستم نیست، درحالی‌که داده‌های پرت به دلیل خطای انسانی، خطای ماشین (ابزار) و به‌طورکلی وجود خطا در سیستم رخ می‌دهند (آلبانیس و بچلور[2]،۲۰۰۷). پدیده‌های استثنایی حجم کمی از داده-های موجود در پایگاه داده را تشکیل می‌دهند. دانش نهفته در رفتار این پدیده‌ها بسیار ارزشمند است. آن‌گونه که در این تحقیق داده‌ها طبقه‌بندی شده‌اند داده‌های استثنایی داده‌های مثبت غیرقابل‌انتظار و داده‌های پرت داده‌های منفی غیرقابل‌انتظار می‌باشند.

شناسایی عناصر و قواعد استثنایی بسیار دشوارتر از شناسایی پدیده‌های معمول است؛ زیرا پدیده‌های استثنایی به تعداد دفعات کمتری از رفتارهای عادی رخ می‌دهند و قابلیت پیش‌بینی بروز آن‌ها از پیچیدگی‌های خاص برخوردار است (نقی[3]،۲۰۰۹). داده‌های استثنایی تأثیر مهمی بر تنزل عملکرد مدل‌های یادگیری دارند؛ زیرا تکنیک‌های کاوش الگوهای استاندارد، به‌منظور کشف و استخراج دانش عمومی نهفته در یک پایگاه داده طراحی شده‌اند و مجموعه داده‌های مشابه را به‌منظور کشف قوانین مربوطه مورد توجه قرار می‌دهند. هنگام مواجهه با داده‌های استثنایی، قوانینی که توسط آن‌ها دسته‌بندی و پیش‌بینی دسته‌های کوچک انجام می‌شود، بسیار کمتر و ضعیف‌تر از قوانینی هستند که دسته‌های اصلی را پیش‌بینی می‌نمایند (جاپکوویچ[4]، ۲۰۰٤). به همین علت نمونه‌های متعلق به طبقه‌های کوچک بیشتر از نمونه‌های متعلق به دسته‌های اصلی

1. Chawla
2. Albanis & Batchelor
3. Nagi
4. Japkowicz

به اشتباه دسته‌بندی می‌شوند (جوشی[1]،۲۰۰۲). از سوی دیگر ماهیت غیرمتوازن داده، عدم ترسیم آستانه‌های دقیق، نبود سنجه ارزیابی مناسب، عدم وجود اطلاعات کافی از دیگر چالش‌های موجود در مسئله یادگیری داده‌های استثنایی است (ویس[2]، ۲۰۰٤)؛ بنابراین تدوین تکنیک یادگیری که بتواند برای تمامی انواع داده و در هر شرایطی به‌خوبی عمل نماید اهمیت زیادی دارد.

هدف اصلی تحقیق پیش روی، کشف و یادگیری از رخدادهای استثنایی از طریق شناسایی داده‌های استثنایی موجود در پایگاه داده و استخراج دانش پنهان این پدیده‌هاست. بدین منظور، نیازمند زبانی برای توضیح ساختار و رفتار سهام استثنایی با استفاده از مجموعه‌ای از اصول و قوانین برای کشف استثنائات هستیم.

پیشینه تحقیق

مطالعات پیشین در حوزه شناسایی و یادگیری داده‌های استثنایی را می‌توان در سه حوزه کشف استثنائات بر اساس نوع داده، بهبود کارایی الگوریتم‌های موجود در مواجهه با داده‌های نامتوازن و مرور و دسته‌بندی پژوهش‌های پیشین طبقه‌بندی نمود. به‌منظور کشف و شناسایی پدیده‌های استثنایی اغلب از تکنیک‌های دسته‌بندی (کیم و شون[3]، ۲۰۱۲ و آلبانیس و بچلور، ۲۰۰۷)، روش‌های آماری (کلارک[4]،۲۰۱٤) و نمونه‌سازی (گانگ[5]، ۲۰۱۰) استفاده شده است. به‌عنوان‌مثال، یوفنگ کو (۲۰۰٦) الگوهای غیر نرمال موجود در پایگاه داده را بر اساس مقیاس فاصله از مراکز دسته شناسایی نمود. وی ماهیت داده‌های غیر نرمال و داده‌های پرت را یکسان انگاشته و رویکرد سامانمند برای کشف مناطق پرت در سری زمانی ارائه کرده است. کو و همکارانش[6] (۲۰۰۸) با به‌کارگیری ابزار آماری به کاوش الگوی فعالیت‌های هدف، فعالیت‌های مغایر با هدف و فعالیت‌های با تأثیر معکوس پرداخته‌اند، درحالی‌که ساختار داده غیرمتوازن در نظر گرفته شده است. بورز[7] و همکارانش (۲۰۰۹) موضوع از دست دادن مشتری را به‌عنوان یک رخداد استثنایی در صنعت خدمات مورد بررسی قرار دادند که هدف آن ارتقا

1. Joshi
2. Weiss
3. Kim & Sohn
4. Clark
5. Gong
6. Kou
7. Burez

کارایی روش‌های نمونه‌سازی با به‌کارگیری سنجه‌های ارزیابی مناسب‌تر است. چن[1] و همکاران (۲۰۰۸) از رویکرد داده‌کاوی بر اساس دانه‌های اطلاعاتی برای کشف دانش از داده‌های نامتوازن استفاده نموده است. این روش باعث به حداقل رساندن دخالت انسان در پردازش اطلاعات شده و دانش لازم را از دانه‌های اطلاعاتی جمع‌آوری می‌کند. گارسیا[2] و همکارانش (۲۰۱۲) کارایی روش‌های پیش‌پردازش را در مواجهه با سطوح مختلف عدم توازن مورد بررسی قرار داده و به این نتیجه رسیدند که در مواجهه با داده‌های غیر متوازن مطلق، نمونه‌سازی افزایشی دسته کوچک بدتر از نمونه‌سازی کاهشی عمل می‌کند. همچنین هو[3] و همکارانش (۲۰۰۹) و دونگ[4] و همکارانش (۲۰۰۵) به کشف رفتارهای پرخطر افراد در خانه‌های هوشمند با استفاده از روش زنجیره مارکوف پرداختند. همچنین مطالعه روند تاریخی پیشرفت سامانه‌های یادگیری از داده‌ها و حالات نامتوازن توسط چاولا و همکارانش (۲۰۰٤) و ویس (۲۰۰٤) بررسی شده است. ویس مسائل و مشکلات کاویدن داده‌ها و حالات نادر دریک مجموعه بزرگ از داده را بررسی نموده و به‌تفصیل به روش‌های مقابله با این مشکلات پرداخته است.

با بررسی تحقیقات پیشین درمی‌یابیم که هیچ‌یک از پژوهش‌های موردمطالعه آستانه مشخصی برای تفکیک داده‌های پرت و داده‌های استثنایی ارائه ننموده‌اند (بوشز،۲۰۰۹). در حالتی که داده‌ها از نوع چندرسانه‌ای هستند فقط به شناسایی استثنائات پرداخته شده و در بقیه موارد داده‌های پرت و استثنایی توأمان شناسایی شده‌اند (ژیانگ و گونگ[5]، ۲۰۰۸). با نظر به اهمیت دانش نهفته در داده‌های استثنایی و زایل بودن داده‌های پرت، یکی از اهداف تحقیق پیش روی تفکیک داده‌های استثنایی از داده‌های پرت است. این مهم با به‌کارگیری تابع آنتروپی و جداسازی استثنائات بر اساس محتوای اطلاعاتی محقق می‌شود. همچنین به‌کارگیری رویکرد یادگیری پایین به بالا کاستی‌های الگوریتم-های یادگیری معمول مانند درخت تصمیم را بهبود می‌بخشد.

---

1. Chen
2. García
3. Hu
4. Duong
5. Xiang & Gong

اهداف پژوهش

پژوهش حاضر با ارائه مدل جدیدی به شناسایی استثنائات و یادگیری رفتار آن‌ها می‌پردازد. نخست با به‌کارگیری شاخص استخراج اطلاعات (آنتروپی[1])، داده‌های استثنایی را از داده‌های نرمال[2] تفکیک نموده و سپس با به‌کارگیری الگوریتم پیشنهادی E-RISE بر اساس رویکرد یادگیری پایین به بالا[3] قوانین بروز رفتار استثنایی استخراج می‌گردد. داده‌های استثنایی با استفاده از تابع آنتروپی رنی شناسایی می‌گردد. آنتروپی رنی با اختصاص وزن بیشتر به داده‌های استثنایی باعث شاخص شدن این نوع پدیده‌ها می‌شود. قوانین بروز رفتار استثنایی توسط الگوریتم پیشنهادی RISE ارتقا یافته استخراج گردیده که برای ساخت سیستم خبره‌ای به‌منظور شناسایی استثنائات جدید بکار گرفته می‌شود. در پژوهش حاضر با به‌کارگیری مدل پیشنهادی، سهام استثنایی موجود در بازار سهام تهران شناسایی شده و الگوهای بروز رفتار سهام استثنایی کشف می‌گردد. به‌کارگیری روش پیشنهادی موجبات تحقق اهداف زیر فراهم می‌آورد:

۱. ارائه مدل جدیدی برای کشف سهام استثنایی بر اساس نظریه اطلاعات[4] و استفاده از رویکرد یادگیری پایین به بالا برای ارائه الگوریتم RISE ارتقا یافته برای شناسایی دانش پدیده‌های استثنایی.

۲. به‌کارگیری مدل پیشنهادی برای شناسایی سهام استثنایی و یادگیری قوانین رفتاری سهام استثنایی.

۳. طراحی و به‌کارگیری سیستم خبره به‌منظور استفاده از دانش استخراج‌شده برای کشف استثنائات جدید.

٤. ارتقا نظریه استثنائات

چارچوب تحقیق حاضر به‌صورت زیر است. در بخش ٤ در قالب روش‌شناسی پژوهش به بیان نظریه استثنائات پرداخته و سپس مفهوم آنتروپی و چگونگی شناسایی داده‌های استثنایی با به‌کارگیری آنتروپی رنی عنوان می‌شود. دانش رفتار سهام استثنایی با به‌کارگیری رویکرد پایین به بالا و بر اساس روش پیشنهادی RISE ارتقا یافته استخراج می‌گردد. در بخش ۵ مدل پیشنهادی برای کشف سهام استثنایی و استخراج

---

1. Entropy
2. Normal data
3. Bottom-Up learning approach
4. Abnormality Theory

قوانین بروز رفتار استثنایی برای سهام موجود در بازار بورس تهران بکارگرفته شده و درنهایت نتایج حاصل از پژوهش ارائه می‌گردد.

روش‌شناسی پژوهش

منطق یادگیری از استثنائات یک مسئله قابل‌توجه در حوزه یادگیری ماشین است. در پژوهش حاضر مدلی بر اساس رویکرد تلفیقی تئوری استثنائات و تئوری اطلاعات ارائه شده است تا پدیده‌های استثنایی را کشف نموده و الگوهای رفتاری پنهان آن‌ها را شناسایی نماید. تئوری اطلاعات با هدف کشف استثنائات به‌عنوان راهبردی برای اندازه‌گیری میزان بی‌نظمی‌های مجموعه داده به کارگرفته می‌شود، سپس دانش پدیده‌های استثنایی و نرمال توسط الگوریتم یادگیری E-RISE کشف می‌گردد. چرخه شناسایی استثنائات، یادگیری رفتار آن‌ها و به‌کارگیری قوانین شناسایی‌شده با به‌کارگیری تئوری استثنائات و طراحی سیستم خبره تشخیص استثنائات تکمیل می‌گردد.

نظریه استثنائات

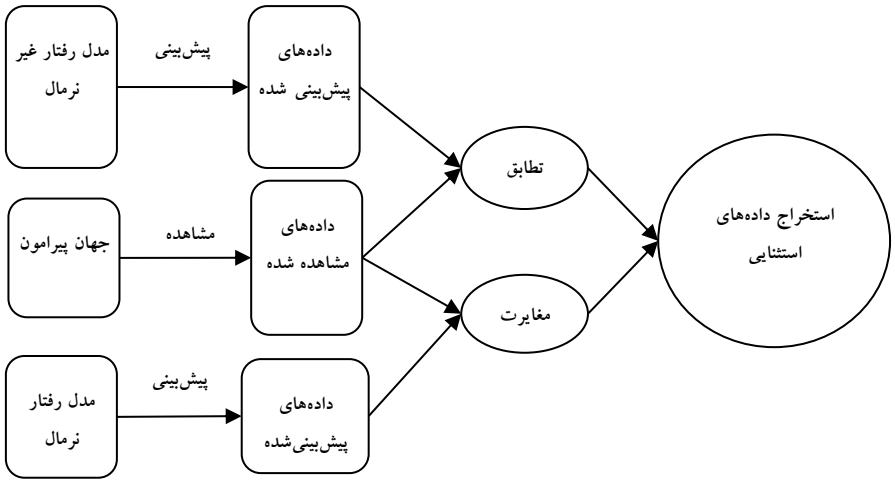
رویکردهای متفاوتی به مسئله استثنائات در حیطه‌های متفاوت علمی و عملی وجود دارد که از آن‌ها می‌توان به استثنائات موضوعی، استثنائات آماری، استثنائات ژنتیکی، استثنائات بیولوژیک، و رویکردهای تئوریک اشاره نمود. نخستین بار هافمن[1] (۱۹۷۷) تئوری استثنائات را در علم ژنتیک مطرح نمود. پس از ظهور تئوری استثنائات در علم ژنتیک مک کارتی در سال ۱۹۸۰ مفهوم تئوری استثنائات را در استنتاج مطرح نمود. مک کارتی مثال معروفی را برای معرفی استثنائات بیان نمود. این قانون خاص را در نظر بگیرید «پرندگان به‌صورت معمولی می‌توانند پرواز کنند» اگر x یک پرنده خاص باشد Ab(x) به این معناست که "x یک مورد استثنایی از جامعه پرندگانی هست که نمی‌توانند بپرند مانند پنگوئن؛ بنابراین مک کارتی با به‌کارگیری قوانین علی و معلولی به شناسایی استثنائات پرداخت. پس از رویکرد مک کارتی، رویکردهای تئوریک به مسئله شناخت استثنائات مطرح گردید. رویکردهای تئوریک به استثنائات بر اساس یک تئوری که توسط شخصی ایجاد و یا توسعه داده شده، آغاز می‌گردد. اگر حیطه نرمال را بتوان برای مسئله تعریف نمود آنگاه استثنائات به‌عنوان شکست در توسعه این تئوری

1. Hoffman

در نظر گرفته می‌شود.

تئوری مفهومی که برای یافتن استثنائات به کار گرفته می‌شود، بسته به نوع دانش موجود عمل می‌کند. درواقع استثنائات مرتبط با تفسیر یافته‌های مشاهده‌شده در محتوای حوزه مسئله است. نقطه شروع مناسب برای توضیح استثنائات در مدل مفهومی انواع مختلف دانشی است که در کاربردهای مختلف استثنائات نقش دارد. دانش ضمنی موجود در سیستم یافتن استثنائات ممکن است بر پایه توضیحی از ساختار نرمال و رفتار وظیفه‌ای سیستم و یا بیان رفتار غیر نرمال سیستم باشد. یافته‌های جمع‌آوری‌شده‌ای که با رفتار نرمال سیستم مطابقت دارند "یافته نرمال" و در غیر این صورت "یافته غیر نرمال" گفته می‌شود. بر اساس انواع دانش موجود و یافته‌های مشاهده شده، دو دسته تئوری شکل می‌گیرد: تئوری انحراف از ساختار و رفتار نرمال و تئوری انطباق با رفتار غیر نرمال. تئوری انحراف از ساختار و رفتار نرمال تئوری نخستین بار توسط ری ریترز[1] (۱۹۸۷) به‌عنوان چارچوب منطقی برای کشف عارضه‌های سامانه‌های فیزیکی با استفاده از مدل ساختار و رفتار نرمال سیستم مطرح گردید. بر اساس این تئوری، استثنائات بر اساس مقایسه داده‌های مشاهده‌شده با ساختار و رفتار نرمال سیستم و مغایرت با آن کشف می‌شوند. کشف استثنائات بر اساس تطابق با رفتار غیر نرمال با در نظر گرفتن دانش رفتارهای استثنایی سیستم عمل می‌کند، به‌نحوی‌که به شبیه‌سازی رفتار غیر نرمال سیستم می‌پردازد. به‌منظور ارتقا دقت در کشف استثنائات چارچوب تئوریک تلفیقی جدیدی بر اساس تئوری کشف استثنائات بر اساس سازگاری و تئوری کشف استثنائات بر اساس تطابق با رفتار غیر نرمال به‌صورت زیر (نمایه ۱) پیشنهاد می‌گردد. استثنائات بر مبنای میزان سازگاری داده‌های مشاهده‌شده با مدل رفتار غیر نرمال و یا مغایرت داده‌های مشاهده‌شده با رفتار نرمال سیستم کشف می‌گردند.

1. Reiter



شکل ۱. مدل پیشنهادی برای شناسایی داده‌های استثنایی بر اساس تئوری استثنائات

آنتروپی

آنتروپی یکی از مفاهیم اصلی در تئوری اطلاعات است که به‌منظور ارزیابی میزان یکنواختی مجموعه‌ای از داده‌ها بکار می‌رود (ستیوهادی و همکاران، ۲۰۱٤). آنتروپی با مفهوم تصادفی بودن پدیده‌ها در ارتباط است و به‌عنوان راهبردی برای اندازه‌گیری میزان عدم قطعیت، بی‌نظمی و تصادفی بودن یک پیشامد استفاده می‌گردد. اندازه آنتروپی با افزایش میزان همگنی داده کاهش می‌یابد (لی و رو،۲۰۱۲ و کاور و توماس، ۱۹۹۱). در پژوهش حاضر از تابع آنتروپی به‌عنوان ابزاری برای سنجش میزان استثنایی بودن داده‌ها استفاده شده است. بر اساس روش پیشنهادی میزان آنتروپی یک رکورد نشان‌دهنده میزان نفع اطلاعات موجود در آن سطر از داده است. هر سطر از پایگاه داده یک دسته منفرد انگاشته می‌شود و تابع آنتروپی برای هر رکورد محاسبه می‌گردد. هر اندازه تمایل تابع آنتروپی محاسبه‌شده به میانگین آنتروپی داده‌های موجود در پایگاه داده نزدیک‌تر باشد، یعنی داده مربوط به دسته نرمال است و رفتار آن از رفتار نرمال سیستم تبعیت می‌کند (قمر،۲۰۱۱). داده‌هایی که آنتروپی آن‌ها تفاوت معناداری از متوسط آنتروپی داده‌های موجود در پایگاه داده دارد غیر نرمال است. این دسته از داده-ها رفتاری متفاوت از رفتار معمول سیستم از خود بروز می‌دهند.

اگر X یک متغیر تصادفی باشد که یکی از مقادیر x_1 و...و x_2 و x_n را با احتمالات p_1 و...و p_2 و p_n به خود می‌گیرد، $H(x) = -\sum_{i=1}^{n} p_i \log(p_i)$ نشان‌دهنده میزان آنتروپی حاصل از آن است. $H(X)$ نشان‌دهنده میزان عدم قطعیت برای بروز حالت مشخص X است. آنتروپی با تعریف فوق آنتروپی شانون نامیده می‌شود. ازآنجایی‌که هدف ما شناسایی داده‌های استثنایی در پایگاه داده است و تابع آنتروپی رنی برای داده‌هایی با احتمال وقوع کمتر وزن بیشتری را لحاظ می‌کند، از تابع آنتروپی رنی به‌عنوان ابزاری برای استخراج میزان اطلاعات موجود در هر رکورد و تشخیص داده‌های استثنایی از داده‌های نرمال استفاده می‌نماییم؛ زیرا تابع آنتروپی رنی تمایز بین داده‌های نرمال و غیر نرمال را آشکارتر می‌گرداند.

$$H(x) = \frac{1}{1-\propto} \log(\sum p_i^{\propto}) \qquad (۱)$$

تابع آنتروپی رنی با $\propto = 2$ را به کار می‌گیریم. هراندازه $\propto$ کوچک‌تر باشد توانایی تشخیص داده‌های استثنایی آسان‌تر است. بدین ترتیب استفاده از توان ۲ برای احتمالات کوچک باعث بزرگ‌تر شدن مقدار تابع $H(x)$ و وزن دهی بیشتر به آن می‌شود.

$$H(x) = -\log(\sum p_i^2) \qquad (۲)$$

برای محاسبه آنتروپی هر رکورد از رابطه زیر استفاده می‌شود:

$$H(x,y) = H(y) + H(xIy) \qquad (۳)$$

آنتروپی داده‌های غیر نرمال دارای انحراف زیادی نسبت به متوسط تابع آنتروپی دیگر داده‌ها است. چون $H(y)$ به ازای مقادیر کوچک y بزرگ‌تر است بنابراین آنتروپی داده‌های غیر نرمال فاصله بیشتری از بقیه داده‌ها می‌گیرد.

استخراج قوانین

پس از شناسایی داده‌های استثنایی و تفکیک آن از داده‌های نرمال، بایستی دانش رفتار پدیده‌های نرمال و استثنایی کشف شود. الگوریتم‌های یادگیری سنتی مانند درخت تصمیم بر پایه یک عمل جستجوی حریصانه از رویکرد یادگیری بالا به پایین پیروی می‌کنند. این الگوریتم‌ها اغلب در شناسایی قوانین عمومی موجود در پایگاه داده بهتر عمل می‌نمایند. بدین ترتیب که یادگیری با استفاده از مؤثرترین متغیر و ایجاد یک قانون عمومی آغاز و با در نظرگیری یک‌به‌یک متغیرهای مهم با استفاده از ساختاری که به‌طور طبیعی در فضای فرضیه‌ها رخ می‌دهد، صورت می‌پذیرد. این قوانین داده‌های

متعلق به دسته‌های اصلی را پوشش داده و از داده‌های استثنایی صرف‌نظر می‌نمایند. لذا الگوریتم‌های یادگیری سنتی عملکرد مناسبی در رفتار استثنائات ندارند. به‌منظور غلبه بر کاستی‌ها و مشکلات این الگوریتم‌ها در یادگیری از استثنائات، روش جدیدی مبتنی بر رویکرد یادگیری پایین به بالا (کالیف،۲۰۰۳) با نام الگوریتم RISE ارتقا یافته برای کشف دانش استثنائات پیشنهاد شده است. در این رویکرد توجه و تمایل به ایجاد قوانین خاص با دقت بالا است.

بر اساس الگوریتم RISE ارتقا یافته، ابتدا داده‌ها در دو دسته نرمال و استثنایی قرار می‌گیرند. سپس داده‌های نرمال با استفاده از روش‌های خوشه‌بندی و بر اساس میزان شباهتشان، به بیشترین تعداد خوشه ممکن تقسیم می‌شوند. فرایند خوشه‌بندی سبب افزایش دقت قوانین ایجادشده می‌گردد. استخراج قوانین با انتخاب تصادفی درصدی از داده‌های موجود در هر خوشه آغاز می‌شود و منطبق با هر نقطه داده انتخابی یک قانون ساخته می‌شود. سپس قوانین با توجه به نقاط داده موجود در همسایگی آن و با توجه به معیارهای دقت و support عمومی می‌شوند، به‌طوری‌که داده‌های موجود در یک همسایگی معین وجود دارد پوشش داده می‌شوند. عمومی‌سازی با توجه به تابع فاصله اقلیدسی صورت می‌گیرد و با حذف محدودیت روی متغیرها یا باز کردن بازه‌ها صورت می‌گیرد. مدلRISE ارتقا یافته یک مدل یادگیری خاص به عام برای یادگیری رفتار استثنائات است.

جدول ۱. مراحل استخراج قوانین بر اساس روش پیشنهادی یادگیری استثنائات بر اساس نقاط داده

فاز عمومی‌سازی	فاز ایجاد قوانین
اگر $R = (a_1, a_2, \dots, a_m, C_R)$ یک رول و $N = (e_1, e_2, \dots, e_m, C_n)$ یک نمونه نزدیک باشد: function generalization (R,N) $a_i$is either true ,$x_i = r_i$ or $r_{i,lower} \le x_i \le r_{i,upper}$ For attribute i-th IF $a_i = True$then do nothing else if $e_i > r_{i,upper}$then $e_i = r_{i,upper}$ else if $e_i < r_{i,lower}$ then $e_i = r_{i,lower}$	ES is the training set SS= select α % of ES randomly Let RS be SS For each rule R in RS N= the nearest neighborhoods E to R by $d < d^*$ let $\acute{R}$ = Generalization (R,N) Let $\acute{RS} = RS$ with R replaced R with $\acute{R}$ if Acc($\acute{RS}$)>= $\beta\%\ Acc(RS)$ Then replace RS by $\acute{RS}$ if $\acute{R}$ is identical to another rule in RS then delete $\acute{R}$ from RS Until $Acc(RS) \ge \gamma$ Return RS

از نقاط قوت روش پیشنهادی یادگیری، شناسایی استثنائات و قرار دادن آن‌ها در یک دسته جداگانه است. این کار باعث افزایش دقت یادگیری رفتار استثنائات می‌گردد و کاستی‌های مدل‌های معمول یادگیری را برطرف خواهد نمود. همچنین نمایه ۲ فرایند کشف و یادگیری از رفتار استثنائات را نشان می‌دهد. چرخه کشف، یادگیری و به‌کارگیری قوانین با طراحی سیستم خبره‌ای تکمیل خواهد شد که وظیفه کشف استثنائات جدید را بر عهده دارد. استثنائات جدید با استفاده از قوانین ایجادشده در چارچوب تئوری استثنائات شناسایی می‌شوند.

همان‌گونه که در بخش‌های پیشین مطرح شد رویکرد پیشنهادی جدیدی برای شناسایی سهام استثنایی و یادگیری از آن‌ها مطرح گردیده است که در ادامه به‌تفصیل به گام‌های آن می‌پردازیم:

گام ۱ـ محاسبه آنتروپی همه سطرهای داده

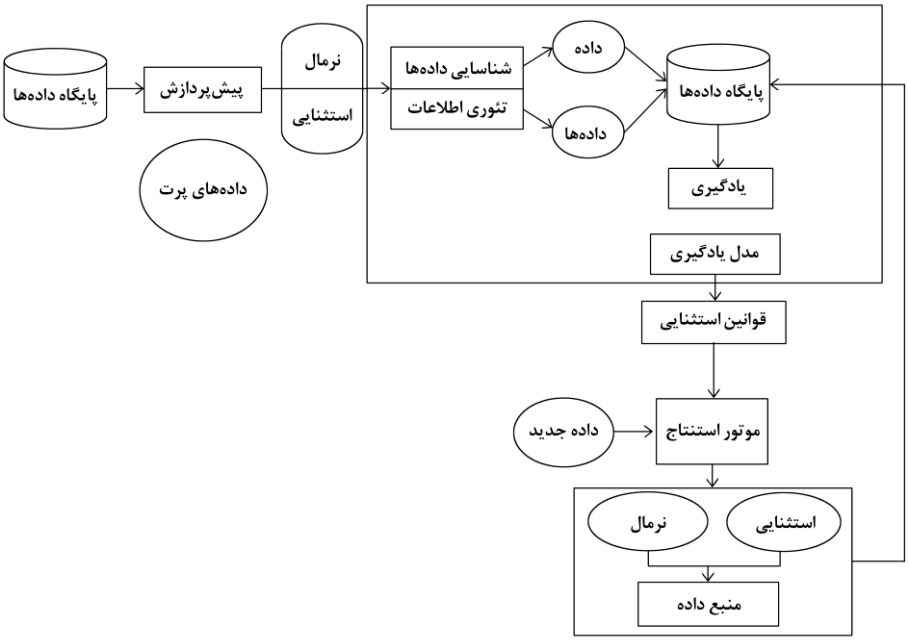
گام ۲ـ شناسایی داده‌های استثنایی ـ داده‌هایی که آنتروپی آن‌ها تفاوت معناداری از میانگین آنتروپی داده‌ها دارد.

گام ۳ـ دسته‌بندی داده‌ها ـ ابتدا داده‌های نرمال و غیر نرمال شناسایی‌شده در گام دو را خوشه‌بندی نموده و سپس داده‌های نرمال را مجدداً به‌گونه‌ای دسته‌بندی می‌شود تا داده‌های مشابه در یک دسته قرار گیرند.

گام ٤ـ درصدی (α %) از داده‌های موجود در هر دسته ایجادشده (گام ۳) را به‌تصادف انتخاب کرده و به ازای هر داده موجود در نمونه تصادفی انتخاب‌شده یک قانون ایجاد می‌شود .

گام ۵ـ قوانین ایجادشده با در نظرگیری داده‌هایی که نزدیک‌ترین فاصله را با آن‌ها دارند و تاکنون توسط قانون دیگری پوشش داده نشده‌اند، عمومی‌تر می‌شوند. این کار توسط حذف شرایط یا باز کردن بازه‌ها برای متغیرهای عددی انجام می‌گیرد. به‌منظور سنجش کارایی فرایند عمومی‌سازی در هر مرحله از شاخص g-means و support استفاده می‌شود. تغییرات در صورتی لحاظ می‌شوند که شاخص g-means بدتر نشده و قانون ایجادشده درصد مشخصی از داده‌های موجود در آن دسته را بپوشاند.

گام ۷- طراحی سیستم خبره برای شناسایی استثنائات جدید با استفاده از چارچوب پیشنهادی تئوری استثنائات. استثنائات جدید بر اساس قوانین رفتاری نرمال و استثنایی کشف می‌شود. بدین ترتیب، داده‌هایی که با قوانین رفتار نرمال مطابقت ندارند یا منطبق با قوانین رفتار استثنایی عمل می‌نمایند، به‌عنوان پدیده‌های استثنایی تلقی می‌شوند.



شکل ۲. متدولوژی پیشنهادی بر پایه رویکرد تلفیقی تئوری استثنائات و اطلاعات

یافته‌های پژوهش

هدف از پژوهش حاضر، به‌کارگیری تئوری اطلاعات برای کشف پدیده‌های استثنایی و شناسایی الگوهای رفتاری استثنائات است. حیطه عملی موردبررسی، رصد رفتار متفاوت سهام مختلف به‌منظور شناسایی و تشکیل سبد سهام استثنایی است. نخستین گام پژوهش، شناسایی و انتخاب متغیرهای دخیل برای تبیین وضعیت سهام و انتخاب سبد سهام استثنایی است. رفتار متغیرها اساس تشخیص رفتار سیستم است. متغیرها بایستی به‌گونه‌ای انتخاب گردند که نسبت به استثنائات حساس باشند. مطالعات پیشین در حوزه کشف و شناسایی سهام استثنایی، سهام غیر نرمال را معادل سهام با بازده بالا دانسته‌اند. در رویکرد حاضر به مسئله، سهامی استثنایی خوانده می‌شود که منفعت استثنایی را برای سهامدار به همراه داشته باشد.

با هدف تشخیص سهام غیرعادی بر اساس مطالعات پیشین و نظر خبرگان متغیرهای مؤثر در بروز رفتار استثنایی سهم شناسایی شده‌اند. با به‌کارگیری آزمون آماری t از بین متغیرهای موردنظر متغیرهای سود هر سهم، سهام جایزه و افزایش سرمایه شرکت برای

تجزیه رفتار سهام برگزیده شده‌اند. این متغیرها تفاوت معناداری را بین سهام نرمال و استثنایی ایجاد می‌نمایند. در فرایند سرمایه‌گذاری استثنایی، ثروت سهامدار رشد استثنایی دارد که این رشد می‌تواند از طریق کسب سود استثنایی ایجاد گردد. همچنین صدور سهام جایزه با حجم بالا می‌تواند باعث چند برابر شدن ثروت سهامدار گردد.

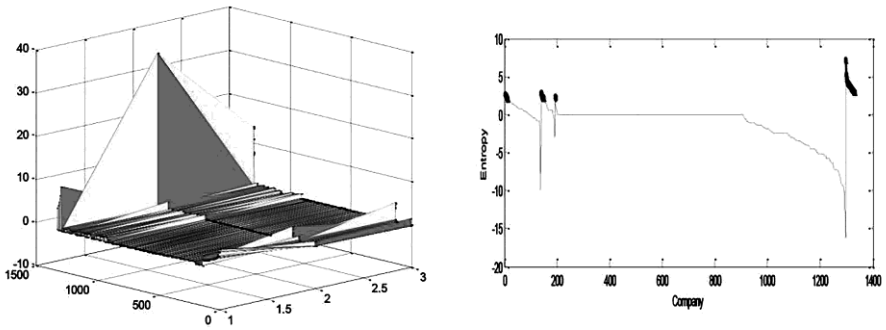
جامعه آماری تحقیق حاضر، اطلاعات ماهیانه پنجاه سهام برتر بازار بورس تهران از ابتدای سال ۱۳۹۲ تا پایان سال ۱۳۹۳ است که در ۱۳۳٤ رکورد گردآوری شده است. به‌منظور تشکیل پروفایل از سهام موردبررسی و کشف استثنائات می‌بایست مجموعه قابل نمایشی از سهام ایجاد گردیده است. فرض کنید $x_{i,j}$ نشان‌دهنده سهم i ام در مشخصه j باشد، درحالی‌که $i = \{1,2, \dots ,1334\}$ نشان‌دهنده ۱۳۳٤ سهم موردبررسی و $j = \{1,2,3\}$ متغیرهای در نظر گرفته شده در شناسایی و یادگیری رفتار استثنایی سهام است. در این صورت M نشانگر پروفایل سهام مورد بررسی است.

$$M = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{bmatrix} \quad (٤)$$

در گام بعد به دلیل اهمیت پاک‌سازی ورودی‌های مدل و تأثیر آلودگی داده‌ها بر تنزل عملکرد فرایند داده‌کاوی، مرحله پیش‌پردازش داده‌ها انجام شده است. بدین‌صورت که داده‌های مفقود با میانگین داده‌های مربوط به سهم موردبررسی جایگزین شده‌اند. به‌منظور نرمالیزه نمودن داده‌ها، از مقدار Z استفاده شده است. لازم به ذکر است داده‌هایی را که خارج از $\pm 3\sigma$ میانگین بوده‌اند به‌عنوان داده‌های پرت در نظر گرفته و از محاسبات خارج گردیده است.

به دلیل رفتار متفاوت داده‌های غیر نرمال از نظر میزان نظم و منفعت اطلاعاتی نسبت به داده‌های نرمال، تابع آنتروپی رنی برای شناسایی و تفکیک داده‌های استثنایی از داده‌های نرمال به‌کاررفته شده است. تابع آنتروپی رنی با ساختار لگاریتمی و پارامتر α عملکرد بهتری در کشف و شناسایی داده‌های غیر نرمال از مجموعه داده در دست را دارد. با بررسی‌های انجام‌شده مشاهده گردید تابع مذکور با $\alpha = 2$ بهترین عملکرد را در تشخیص رفتار استثنایی دارد؛ زیرا رویدادهای غیر نرمال به دلیل احتمال پایین بروز آن‌ها باعث بزرگ‌تر شدن تابع آنتروپی می‌شوند. در پژوهش حاضر سهام استثنایی سهامی هستند که میزان تابع آنتروپی آن‌ها به فاصله $+1/5\sigma$ از میانگین تابع آنتروپی تمامی سهام است. از مجموع ۱۳۳٤ سهم بررسی‌شده، ۳٦ سهم رفتار استثنایی از خود

بروز داده‌اند؛ یعنی ۲/٦ درصد از سهام موردبررسی رفتار استثنایی داشته‌اند. نمایه ۳- الف تفاوت رفتار سهام استثنایی را با ترسیم اطلاعات سهام بر اساس سه متغیر انتخابی نشان می‌دهد. پرواضح است که سهام استثنایی رفتار متفاوت از خود بروز می‌دهند. در نمایه ۳ ـ ب تابع آنتروپی سهام ترسیم شده است. همان‌گونه که در تصویر مشخص است میزان آنتروپی سهام استثنایی با میانگین آنتروپی سایر سهام تفاوت معناداری دارد.



شکل ۳. الف ـ اطلاعات سهام بر اساس سه متغیر انتخابی ب ـ تابع آنتروپی سهام

پس از اجرای مدل درستی و کارایی مدل پیشنهادی مورد بررسی قرار گرفته است. بدین منظور از مقایسه نتایج به‌دست‌آمده با نتایج حاصل از به‌کارگیری روش‌های سنتی برای کشف داده‌های استثنایی، استفاده می‌نماییم. همچنین از شاخص‌های g-means و دقت برای سنجش مطلوبیت مدل پیشنهادی استفاده شده است.

$$(۵)\mathrm{g-means} = \sqrt{\frac{\mathrm{TP}}{\mathrm{TP+FN}} \times \frac{\mathrm{TN}}{\mathrm{TN+FP}}}$$

نتایج نشان داد که مدل پیشنهادی به‌طور قابل‌توجهی عملکرد شناسایی داده‌های استثنایی را بهبود می‌بخشد. نتایج سنجش کارایی روش پیشنهادی در جدول شماره ۲ تلخیص شده است. به‌منظور یادگیری رفتار استثنایی سهام و کسب توانایی در پیش‌بینی سهام استثنایی از رویکرد یادگیری پایین به بالا استفاده شده است. بدین ترتیب که داده‌های موجود در پایگاه داده ابتدا بر اساس رفتار نرمال یا غیر نرمال و سپس با به‌کارگیری روش k-means خوشه‌بندی شده‌اند. این نوع خوشه‌بندی به دلیل استراتژی آن در جداسازی داده‌ها و قابل‌کنترل بودن تعداد خوشه‌ها انتخاب شده است؛ بنابراین سهام در چهار دسته قرار گرفته‌اند که شامل یک خوشه از داده‌های استثنایی و سه خوشه از داده‌های نرمال است. خوشه‌بندی باعث قرارگیری داده‌های نسبتاً مشابه در

یک گروه می‌شود. این‌کار باعث بالا رفتن دقت قوانین استخراج‌شده از آن خوشه می‌گردد. اطلاعات مربوط به خوشه‌ها در جدول ۳ خلاصه شده است.

جدول ۲. سنجش کارایی روش پیشنهادی

شاخص / روش	SVM	CHAID	C5.0	روش پیشنهادی
دقت	۹۷/٤	۹٦/۸	۹۸/۳	۱۰۰
g-means	۰/۹٤٦	۰/۹۳	۰/۹۵	۱

جدول ۳. اطلاعات خوشه‌ها

خوشه	دسته	تعداد	میانگین			انحراف معیار		
			سود سهم	سهام جایزه	افزایش سرمایه	سود سهم	سهام جایزه	افزایش سرمایه
۱	استثنایی	۳٦	۳٤٤۲	۲۱۹/٤	۲۹۱	۳٤۷۳/۲	۲۷۵/۹۷	۱٦۹۷/۸۳
۲	نرمال	۱۱۰۵	۲٦۰	۲/۹	۱٦/۲	۲۸۱	۱۱	۵۸/٤
۳	نرمال	۱۳۸	۲۰۷۵/٦	۰/۹۱	۵/۱٤	۷۸۲,۷	۸/۷۷	۳۱
٤	نرمال	۵۳	۰	۱٤۱/۵۰	۲٤/۳	۰	۵۲/۵	٦۸/٤۳

درصدی (α %) از سهام موجود در هر خوشه به‌تصادف انتخاب و به ازای هر داده، یک قانون منطبق بر آن ساخته شده‌است. قابل‌ذکر است مقادیر متفاوتی برای α در خوشه‌های مختلف لحاظ شده است. جدول ٤ تأثیر تغییر در مقدار α را بر دقت پیش‌بینی و تعداد قانون استخراج‌شده نشان می‌دهد.

جدول ٤. تعیین تعداد بهینه قوانین مستخرج

خوشه	α	α	α	α	α
۱	٪۱۰	٪۵	٪۱۰	٪۱۲	٪۱۵
۲	٪۱۰	٪۷	٪۵	٪۲	٪۲
۳	٪۱۰	٪۷	٪۵	٪۵	٪۲,۵
٤	٪۱۰	٪۵	٪۵	٪۷	٪۷
دقت پیش‌بینی	٪۷۵	٪۷۸	٪۸۸	٪۹٦	٪۱۰۰
تعداد قانون استخراجی	۱۳٤	۹۰	٦٦	۳٤	۳۱

بر اساس بررسی‌های صورت گرفته بهترین مقدار برای α در خوشه‌های ۱ و ۲ و ۳ و ٤ به ترتیب برابر با ۱۵٪، ۲٪، ۲,۵٪ و ۵٪ است. داده‌های خوشه ۱ مربوط به سهام

استثنایی است. ازآنجایی‌که تعداد نقاط داده مربوط به سهام استثنایی اندک (۲/۶٪ از حجم داده) است لذا نیاز به تعداد قانون اولیه بیشتری (۱۵٪ معادل ۵ قانون) برای استخراج دانش پنهان در این‌گونه سهام وجود دارد. در این حالت دقت پیش‌بینی در بیشینه مقدار خود قرار دارد و مسئله از لحاظ پیچیدگی و زمان حل به کمترین میزان خود می‌رسد. در مرحله بعد بر اساس رویکرد یادگیری پایین به بالا و با استفاده از تابع فاصله قوانین اولیه را عمومی‌تر شده است تا جایی که تمامی داده‌ها پوشش داده شود. به‌منظور سنجش فاصله سهام تا قوانین ایجادشده از تابع فاصله اقلیدسی استفاده می‌نماییم. سهامی که نزدیک‌ترین فاصله را با قانون ایجادشده دارند و تابه‌حال توسط هیچ‌یک از قوانین پوشش داده نشده‌اند، به کار گرفته می‌شوند تا با حداکثر دقت و حداقل عمومی‌سازی در قوانین اولیه، قوانین جدید ایجاد شود. این کار توسط حذف شرایط یا باز کردن بازه‌ها برای متغیرهای عددی با حفظ میزان دقت از پیش تعیین‌شده برای قانون ایجادشده انجام می‌گیرد. درصورتی‌که اندازه فاصله کم باشد تعداد تکرارها و زمان عمومی‌سازی افزایش خواهد یافت و درصورتی‌که این فاصله زیاد باشد، دقت قوانین کاهش خواهد یافت. با در نظر گرفتن $k = ۰/۳$ قوانین استخراجی برای تشخیص داده‌های استثنایی ۵ عدد است که با به‌کارگیری قوانین همپوشانی در ۲ قانون به‌صورت زیر خواهد بود:

اگر سود هر سهم$\leq$ ۵۰۰۰ باشد آنگاه سهم استثنایی است.

اگر تعداد سهام جایزه$\leq$ ۳۰۰ باشد آنگاه سهم استثنایی است.

با نظر به دانش رفتاری سهام استثنایی مشخص است که متغیر افزایش سرمایه، نقشی در بروز رفتار استثنایی سهام ندارد. پس از شناسایی قوانین بروز رفتار نرمال و استثنایی سهام، سیستم خبره‌ای بر اساس مدل پیشنهادی تئوری استثنائات طراحی شده است که با ورود اطلاعات سهام جدید، توانایی شناسایی سهام استثنایی را دارد. این سیستم در مواجهه با اطلاعات سهام جدید به دو صورت سهام استثنائی را شناسایی می‌نماید: سهامی که رفتارشان با هیچ الگوی نرمالی مطابقت ندارد و یا سهامی که رفتارشان با یک الگوی استثنایی مطابقت دارد.

به‌منظور سنجش صحت عملکرد مدل پیشنهادی کشف و یادگیری سهام استثنایی، مدل حاضر بر روی داده‌های سه ماهه آخر سال ۱۳۹۳ بکارگرفته شده است. همچنین از خبرگان خواسته‌شده تا سهام استثنایی موجود در این دوره را شناسایی نمایند. نتایج مقایسه خروجی مدل حاضر و نظرات خبرگان بازار بورس نشان‌دهنده تطبیق دانش

استخراج‌شده در مورد سهام استثنایی با نظر خبرگان است.

نتیجه‌گیری

هدف از این تحقیق، طراحی ابزاری کارآمد برای کشف و شناسایی داده‌های استثنایی از میان انبوه داده موجود در پایگاه داده و یادگیری رفتار آن‌ها است. اساس روش پیشنهادی، تدوین مدل کشف و یادگیری استثنائات با به‌کارگیری تئوری اطلاعات و روش یادگیری RISE تعدیل یافته است. ابتدا با استفاده از شاخص استخراج اطلاعات (آنتروپی) به شناسایی داده‌های استثنایی موجود در پایگاه داده پرداخته و سپس با به-کارگیری روش پیشنهادی یادگیری استثنائات بر اساس نقاط قوانین بروز رفتار استثنایی استخراج می‌گردد. به‌منظور سنجش کارایی روش پیشنهادی به کشف و یادگیری سهام استثنایی موجود در بازار بورس ایران پرداخته شده است. اطلاعات مربوط به شرکت‌های حاضر در بازار بورس تهران در ۱۳۳٤ رکورد مورد بررسی قرار گرفت. از مجموع داده‌های موردبررسی ۲/٦٪ استثنایی هستند. از قوانین استخراجی برای پیش‌بینی رفتار سهام جدید و تشکیل پرتفوی بهینه استفاده می‌شود. از نقاط قوت پژوهش حاضر، تفاوت دیدگاه حاکم بر تحقیق با مطالعات موجود، در نحوه کشف استثنائات، به‌کارگیری روش E-RISE برای یادگیری با رویکردی متفاوت، طراحی سیستم خبره برای استفاده از قوانین استخراج‌شده، تعریف متفاوت از سهام استثنایی و بکار گیری ابزارهای کارآمد برای کشف آن‌ها است. هدف ما حداکثر سازی اطمینان به قوانین رفتار استثنائات برای افزایش ثروت سهامداران با شناسایی سهام استثنایی بر مبنای بروز رفتار فراتر از انتظار مثبت و تشکیل پرتفوی استثنایی است. تحقیقات آتی می‌تواند با در نظر گرفتن مجموعه متغیرهای بیشتری که بروز رفتار استثنایی سهام را بیشتر توجیه نماید مانند سود تقسیمی، حق تقدم، درجه نقد شوندگی، نوسانات قیمت و... انجام گیرد. همچنین پیشنهاد می‌شود به‌منظور افزایش دقت تشخیص آستانه رفتار استثنایی سهام از ریاضیات فازی استفاده گردد.

منابع

Albanis G. Batchelor R. Combining heterogeneous classifiers for stock selection, **Intelligent Systems in Accounting, Finance and Management**, vol. 15, no. 1-2, pp. 1-27, 2007.

Burez J. Van den Poel D. Handling class imbalance in customer churn prediction, **Expert Systems with Applications** 36, 4626–4636, 2009

Califf M. E. Mooney R. J. Bottom-Up Relational Learning of Pattern Matching Rules for Information Extraction, **Journal of Machine Learning Research** 4,177-210, 2003.

Cao L. Zhao Y. Zhang C. Mining Impact-Targeted Activity Patternsin Imbalanced Data, **IEEE Transactions on knowledge and data engineering**, Vol. 20, NO. 8, 2008.

Chawla N. V. Japkowicz N. lcz A. K. Editorial: Special Issue on Learning from Imbalanced Data Sets, **Sigkdd Explorations**, 6(1):1–6, 2004.

Chen M. C. Chen L. S. Hsu C. C. Zeng W. R. An information granulation based data mining approach for classifying imbalanced data, **Information Sciences** 178, 3214–3227, 2008.

Clark E. Exploiting stochastic dominance to generate abnormal stock returns, **Journal of Financial Markets** 20, 20–38, 2014.

Cover T. M. Thomas J. A. Entropy, Relative Entropy and Mutual Information; Elements of Information Theory, **ISBN** 0-471-06259-6-pp: 12-49, 1991.

Duong T. V. Bui H. H. Phung D. Q. Venkatesh S. Activity Recognition and Abnormality Detection with the Switching Hidden Semi-Markov Model, **IEEE Computer Society Conference on Computer Vision and Pattern Recognition** (CVPR'05), 2005.

García V. Sánchez J.S. Mollineda R.A. On the effectiveness of preprocessing methods when dealing with different levels of class imbalance, **Knowledge-Based Systems** 25, 13–21, 2012.

Gong R.S. A Segmentation and Re-balancing Approach for Classification of Imbalanced Data, PHD theses, **University of Cincinnati**, 2010.

Hoffman M. L. Moral internalization: Current theory and research, In L. Berkowitz (Ed.), **Advances in experimental social psychology**10, 85-133, 1977.

Hu D. H. Zhang X. X. Yin J. Zheng V. W. Yang Q. Abnormal Activity Recognition Based on HDP-HMM Models, **the Twenty-First International Joint Conference on Artificial Intelligence**, 2009.

Japkowicz, N., The class imbalance problem: Significance and strategies, the international conference on artificial intelligence: **Special track on inductive learning**, 2000.

Joshi M. V, Learning Classifier Models for Predicting Rare Phenomena, PhD thesis, **University of Minnesota**, Twin Cites, Minnesota, USA, 2002.

Kim Y. Sohn S.Y. Stock fraud detection using peer group analysis, **Expert Svstems with Applications** 39, 8986–8992, 2012.

Kou Y, Abnormal Pattern Recognition in Spatial Data, PHD theses, **Faculty of Virginia Polytechnic Institute and State University**, 2006.

Li X. Rao F. Outlier Detection Using the Information Entropy of

Neighborhood Rough Sets, **Journal of Information & Computational Science**, 3339–3350, 2012.

McCarthy J. Applications of circumscription to formalizing common-sense knowledge,**Artificial Intelligence** 28, 89-116, 1986.

Nagi J. An intelligent system for detection of non-technical losses in Tanaga National Berhad (TNB) Malaysia low voltage distribution network, PhD Thesis, **Tenaga national university**,2009.

QamarU.Automated Entropy Value Frequency (AEVF) Algorithm for OutlierDetection in Categorical Data, **Recent Advances in Knowledge Engineering and Systems Science**,28-35, 2011.

Reiter R. A Theory of Diagnosis from First Principles, **Artificial Intelligence** 32, 57-95, 1987.

Setyohadi D. B. Abu Bakar A. Othman Z.A. Rough K-means Outlier Factor Based on Entropy Computation, **Research Journal of Applied Sciences, Engineering and Technology** 8(3): 398-409, 2014.

Weiss G. Mining with rarity: A unifying framework. SIGKDD Explorations Special Issue on Learning from Imbalanced Datasets,6(1):7–19, 2004.

Xiang T. Gong S. Video Behavior Profiling for Anomaly Detection. **IEEE Trans. on Pattern Analysis and Machine Intelligence** 30(5), 893–908, 2008.