ارائهٔ رویکرد نوین یادگیری ماشین برای شناسایی و تجزیه و تحلیل دانش پدیده‌های استثنایی

الهه حاجی‌گل یزدی*، مسعود عابسی، محمد باقر فخرزاد و حسن حسینی نسب
دانشکده فنی مهندسی، گروه مهندسی صنایع، دانشگاه یزد، یزد، ایران

چکیده

کشف پدیده‌های استثنایی پنهان در حجم انبوهی از رکوردهای موجود در پایگاه داده و استخراج دانش آن‌ها در این مطالعه مورد بررسی قرار گرفته است. پدیده‌های استثنایی به‌ندرت رخ می‌دهد و در حجم انبوهی از داده‌های عادی پنهان‌اند. دست‌یابی به دانش رفتاری این پدیده‌ها، ارزشمند و جذاب است. روش‌های موجود یادگیری، در هنگام پاک‌سازی پایگاه داده اغلب پدیده‌های استثنایی را به‌عنوان داده‌های پرت شناسایی کرده و از محاسبات خارج می‌کند و یا اینکه به‌دلیل تمایل به کلّیت، قابلیت شناسایی و دسته‌بندی درست این پدیده‌ها را ندارند. به همین دلیل، ایجاد چارچوبی کارآمد برای کشف دانش و یادگیری رفتار پدیده‌های استثنایی معدود که در میان انبوه رکوردهای یک پایگاه داده مخفی هستند، حائز اهمیت است. در این پژوهش، با به‌کارگیری تئوری استثنائات و تئوری‌های اطلاعات و دانه‌بندی اطلاعات نسبت به استخراج دانش رفتار پدیده‌های استثنایی اقدام شده است. کارآیی روش پیشنهادی با درنظرگرفتن اطلاعات ۳۰ ماههٔ سهام شرکت‌های فعال در بازار اوراق بهادار ایران به‌منظور شناسایی و یادگیری رفتار سهام استثنایی، سنجیده می‌شود.

واژگان کلیدی: داده‌کاوی، پدیده‌های استثنایی، تئوری استثنائات، رویکرد یادگیری پایین به بالا، تئوری اطلاعات.

A Machine Learning Novel Approach for Exceptional Phenomena Knowledge Discovery

Elahe Hajigol Yazdi*, Masoud Abessi, Hasan Hoseini Nasab and Mohammad Bagher Fakhrzad

Department of Industrial Engineering, Yazd University, Yazd, Iran

Abstract

Learning logic of exceptions is a substantial challenge in data mining and knowledge discovery. Exceptional phenomena detection takes place among huge records in a database which contains a large number of normal records and a few of exceptional ones. This is important to promote the confidence to a limited number of exceptional records for effective learning. In this study, a new approach based on the abnormality theory, information and information granulation theories are presented to detect exceptions and recognize their behavioral patterns. The efficiency of the proposed method was determined by using it to detect exceptional stocks from Iran stock market in a 30-month- period and learn their exceptional behavior. The proposed Enhanced RISE algorithm (E RISE) as a bottom-up learning approach was implemented to extract the knowledge of normal and exceptional behavior. The extracted knowledge was utilized to design an expert system based on the proposed abnormality theory to predict new exceptions from 6022 stocks. The superior findings show the results of this proposed approach in exceptional phenomena detection, is in accordance with experts' opinions.

* نویسندهٔ عهده‌دار مکاتبات

Keywords: **Data mining, Exceptional phenomena, Abnormality theory, Bottom-Up learning approach, ERISE Algorithm, Information theory.**

۱-مقدمه

منطق یادگیری استثنائات به‌عنوان چالشی جدید در حیطه یادگیری ماشین، زمانی به عرصه ظهور رسید که یادگیری ماشین از یک دانش ابتدایی به یک فناوری کاربردی در دنیای کسب‌وکار ارتقا یافت [15]. مجموعه داده‌های موجود در یک پایگاه داده به دو دسته عادی و غیرعادی تقسیم می‌شوند؛ پدیده‌های عادی از الگوهای اصلی و عمومی موجود در یک پایگاه داده تبعیت می‌کنند. پدیده‌هایی که رفتاری مغایر با این الگوها داشته باشند، غیرعادی تلقّی می‌شوند.

در تعریف پدیده‌های استثنایی، دیدگاه‌های مختلفی موجود است؛ نخستین‌بار *دیویس* در سال۱۹۷۷، پدیده‌های نادر را در قالب مسألهٔ عارضه‌یابی سامانه‌های فیزیکی و براساس انحراف از ساختار و رفتار طبیعی، شناسایی کرد. پس از آن، *هنری کوهن* (۱۹۸۱) "غیرعادی‌بودن" را به‌عنوان "انحراف آماری از یک استاندارد" مطرح کرد که از اشکالات تعریف آماری پدیده‌های استثنایی، وجود نقاط برش اختیاری برای تشخیص رفتار استثنایی است که ممکن است باعث عدم کفایت آن در کشف استثنائات شود؛ سپس *ریترز* و *مک کارتی* شکست‌های یک سامانه فیزیکی را حالات غیر طبیعی سامانه تلقی کرده و با به‌کارگیری قوانین علّی و معلولی، به شناسایی آن پرداختند[20]. در پژوهش پیش‌روی، پدیده‌های استثنایی رکوردهای نادر پنهان در یک پایگاه داده هستند که رفتاری مثبت و متفاوت از الگوهای رفتاری عمومی موجود در آن پایگاه داده، از خود بروز می دهند.

کشف دانش پدیده‌های استثنایی مسألهای چالش‌برانگیز است. یک پایگاه داده بزرگ شامل انبوهی از رکوردهاست که اغلب از الگوهای رفتاری مورد انتظار سامانه تبعیت می‌کنند. در این میان تعداد بسیار معدودی پدیدهٔ استثنایی وجود دارد که رفتاری مثبت و فراتر از انتظار از خود بروز می‌دهند. دست‌یابی به دانش رفتاری این پدیده‌ها، ارزشمند و جذاب است. روش‌های موجود یادگیری، در هنگام پاکسازی پایگاه داده اغلب پدیده‌های استثنایی را به‌عنوان داده‌های پرت شناسایی کرده و از محاسبات خارج می‌کند و یا اینکه به‌دلیل تمایل به کلیت، قابلیت شناسایی و دسته‌بندی درست این پدیده‌ها را ندارند. در این مطالعه تلاش شده است تا با ایجاد چارچوبی کارآمد، الگوهای رفتاری پدیده‌های استثنایی، کشف شود و از دانش فراگرفته شده، برای پیش‌بینی حالات استثنایی جدید، استفاده شود.

۲-پیشینه پژوهش

در سال‌های اخیر، پژوهش‌های گسترده‌ای حول مسأله شناسایی و پیش‌بینی پدیده‌های نادر صورت گرفته است. بسیاری از مطالعات تنها به مقوله شناسایی پدیده‌های نادر موجود در یک پایگاه داده، پرداخته‌اند و از تکنیک‌های مختلفی از جمله خوشه‌بندی، روش‌های آماری، نزدیک‌ترین همسایگی و ... برای شناسایی استثنائات استفاده نموده‌اند[10].

روش‌های متعددی که برای یادگیری رفتار پدیده‌های نادر ارائه شده است، از دو رویکرد یادگیری داخلی و یا خارجی تبعیت می‌کنند. رویکردهای داخلی بر ایجاد الگوریتم‌های جدید یا تعدیل الگوریتم‌های موجود تأکید دارد؛ به‌نحوی‌که بتوان مسأله عدم توازن داده‌ها را مدیریت کرد. رویکردهای خارجی مانند روش‌های نمونه‌سازی با ارائه روش‌های کارآمد پیش‌پردازش داده، تأثیرات منفی عدم توازن مجموعه داده را کاهش می‌دهند[26]،[17]،[7].

روش‌های ارائه‌شده برای حل مشکل الگوریتم‌های دسته‌بندی موجود، به‌طور معمول از رویکردهای داخلی برای بهبود عملکرد آن‌ها استفاده می‌کنند؛ که به‌عنوان مثال می‌توان به روش‌های یادگیری حساس به هزینه ، استفاده از بایاس استنتاجیو سنجه‌های ارزیابی مناسب‌تر، به‌کارگیری روش‌های جستجوی غیر طماع و یادگیری حالات استثنایی اشاره کرد[16]،[12]،[11]،[8]،[5].

روش‌های حساس به هزینه با درنظرگرفتن هزینهٔ خطای دسته‌بندی بیشتر برای پدیده‌های نادر، باعث ایجاد تمایل بیشتر به دسته‌بندی آن‌ها می‌شود. حساسیت به هزینه با تغییر در نسبت داده‌های مثبت به منفی در مجموعه داده-های آموزشی ایجاد می‌شود. این روش به دانش اولیه در مورد برچسب پدیده‌های عادی و استثنایی نیاز دارد. همچنین به-کارگیری ماتریس هزینه و تعیین مقدار هزینه اختیاری برای دسته‌ها باعث می‌شود، قابلیت تعمیم به همه پایگاه‌های داده را نداشته باشد. استثنائات خود از لحاظ ماهیت رفتاری و

اهمیت با هم متفاوتند؛ لذا درنظرگرفتن یک هزینه برای دسته‌بندی اشتباه استثنائات نتیجه دلخواه را ایجاد نمی‌کند.

به‌کارگیری روش‌های جستجوی غیرطماع و به‌طور خاص الگوریتم ژنتیک، به‌دلیل ترسیم فضای مسأله متناسب با یادگیری رفتار استثنایی سامانه، انتخاب تابع مطلوبیت مناسب و تنظیم پارامترهای بهینه برای شناسایی استثنائات دشوار است. عملگر جهش باید با توجه به نیازمندی‌ها و محدودیت‌های مسأله عمل کند و در طول مدت تکامل نیز، معنی‌دار باشد.

در رویکرد "فقط یادگیری حالات استثنایی" به‌عنوان یک تکنیک "دسته‌بندی نیمه‌نظارتی" فقط به مدل‌سازی عناصر استثنایی پرداخته می‌شود که به دانش اولیه در مورد برچسب داده‌های استثنایی احتیاج دارد. همچنین با توجه به تعداد اندک پدیده‌های استثنایی، اطمینان به قواعد شناسایی‌شده در این رویکرد، کم است.

یکی از روش‌های بهبود عملکرد الگوریتم‌های موجود در فراگیری رفتار استثنائات با رویکرد "یادگیری خارجی"، نمونه‌سازی است؛ که سعی در متعادل‌سازی توزیع اولیه پدیده‌های عادی و استثنایی با افزودن نقاط داده جدید و یا حذف نقاط داده موجود دارند. این روش‌ها عملکرد الگوریتم‌های دسته‌بندی سنتی را در یادگیری استثنائات ارتقا بخشند. تلفیق این روش‌ها و سایر روش‌های یادگیری استثنائات، باعث افزایش احتمال شناسایی داده‌های استثنایی و بهبود دقّت یادگیری می‌شود [21]،[7].

چالش‌های دیگری نیز در مسألهٔ یادگیری از پدیده‌های استثنایی وجود دارد که از جمله می‌توان به عدم یکنواختی ماهیت پدیده‌های استثنایی، مشخصه‌های داده، فقدان اطلاعات کافی، عدم ترسیم آستانه‌های دقیق شناسایی استثنائات اشاره کرد[25]. ازاین‌رو لازم است مدلی کارآمد برای کشف رخدادهای استثنایی موجود در پایگاه داده و استخراج دانش رفتاری آن‌ها طراحی شود.

۳- اهداف پژوهش

در مطالعات انجام‌شده، تمام روش‌های پیشنهادی برای یادگیری رفتار استثنائات از رویکرد یادگیری از بالا به پایین، برای یادگیری استفاده کرده‌اند. این شیوه به‌دلیل شکست فضای مسئله براساس اطلاعات کسب‌شده از متغیرها، باعث ازدست‌رفتن اطلاعات می‌شود. با درنظرگرفتن پژوهش‌های صورت‌گرفته در زمینه کشف استثنائات و یادگیری رفتار آنها [4]،[11]،[25]،[23]، رویکرد نوینی برای حل چالش‌های موجود در مسألهٔ شناسایی داده‌های استثنایی و یادگیری قوانین رفتاری آن‌ها پیشنهاد شده است؛ به‌طوری‌که بتوان به تعداد اندک پدیده‌های استثنایی موجود در یک پایگاه داده برای کشف الگوهای رفتاریشان با اطمینان بالایی اعتماد کرد. در پژوهش حاضر از نمونه‌سازی کاهشی، رویکرد شناسایی استثنائات و قراردادن آن‌ها در دستهٔ جداگانه، رویکرد یادگیری از پایین به بالا و تئوری استثنائات برای یادگیری رفتار استثنائات استفاده شده است؛ به‌طوری‌که قابلیت به‌کارگیری برای هر نوع مجموعه داده -اعم از کمی و یا کیفی- را داشته باشد.

در این مطالعه، با شناسایی صفات و ویژگی‌های مؤثر بر بروز رفتار استثنایی براساس تئوری‌های اطلاعات، دانه‌بندی اطلاعات و استثنائات، پدیده‌های استثنایی کشف شده و در دستهٔ جداگانه‌ای قرار می‌گیرد؛ سپس با استفاده از رویکرد یادگیری از پایین به بالا و براساس الگوریتم پیشنهادی RISE ارتقایافته، رفتار داده‌های عادی و استثنایی، فرا گرفته می‌شود. با استفاده از این رویکرد تمام پدیده‌های موجود در پایگاه داده مورد بررسی قرار گرفته و قوانین رفتاری به‌طور موازی و بر اساس استراتژی پوشش متوالی ساخته می‌شوند که مشکل ازدست‌رفتن اطلاعات را به کمترین میزان خود می‌رساند و منجر به استخراج قوانین رفتاری خاص و با دقت بالا می‌شود. از دانش استخراج‌شده در قالب یک سامانه خبره به‌منظور پیش‌بینی داده‌های استثنایی جدید استفاده می‌شود. به‌کارگیری مدل ارائه‌شده برای شناسایی استثنائات و یادگیری رفتار آن‌ها، برای هر نوع داده اعم از کمی و یا کیفی کاربرد دارد.

روش بالا از تجزیه اطلاعات و ازدست‌رفتن آنها، جلوگیری می‌کند. قواعد عادی و استثنایی استخراج‌شده در قالب چارچوب تلفیقی پیشنهادی تئوری استثنائات، برای طراحی سامانه خبره به‌کار گرفته می‌شود تا چرخه کشف، یادگیری و تبدیل اطلاعات کامل شود.

به‌کارگیری روش پیشنهادی موجبات تحقق اهداف زیر فراهم می‌آورد:

ارائه مدل جدیدی برای کشف استثنائات براساس نظریه اطلاعات و دانه‌بندی اطلاعات و استفاده از رویکرد یادگیری پایین به بالا برای ارائه الگوریتم RISE ارتقایافته برای شناسایی دانش پدیده‌های استثنایی.

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

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

ارتقای نظریه استثنائات

چارچوب پژوهش حاضر بدین صورت است: در بخش ۴ در قالب روش‌شناسی پژوهش به بیان نظریه استثنائات پرداخته و سپس مفهوم نمونه‌سازی کاهشی، آنتروپی و چگونگی شناسایی داده‌های استثنایی با به‌کارگیری آنتروپی رنی عنوان می‌شود. دانش رفتار سهام استثنایی با به‌کارگیری رویکرد پایین به بالا و براساس روش پیشنهادی RISE ارتقایافته استخراج می‌شود. در بخش ۵ مدل پیشنهادی برای کشف سهام استثنایی و استخراج قوانین بروز رفتار استثنایی برای سهام موجود در بازار بورس تهران به‌کارگرفته شده و درنهایت نتایج حاصل از پژوهش ارائه می‌شود.

۴- روش‌شناسی پژوهش

منطق یادگیری از استثنائات یک مسئله قابل توجه در حوزه یادگیری ماشین است. در پژوهش حاضر مدلی بر اساس رویکرد تلفیقی تئوری استثنائات، تئوری اطلاعات و تئوری دانه‌بندی اطلاعات ارائه شده است تا پدیده های استثنایی را کشف کرده و الگوهای رفتاری پنهان آنها را شناسایی کند. ابتدا با هدف کشف استثنائات از تئوری‌های دانه‌بندی اطلاعات برای شکست فضای مسأله و از تئوری اطلاعات به عنوان ابزاری برای اندازه‌گیری میزان بی‌نظمی‌های مجموعه داده استفاده می‌شود؛ سپس دانش پدیده‌های استثنایی و نرمال توسط الگوریتم یادگیری E-RISE کشف می‌شود. چرخه شناسایی استثنائات، یادگیری رفتار آنها و به‌کارگیری قوانین شناسایی‌شده با به‌کارگیری تئوری استثنائات و طراحی سامانه خبره تشخیص استثنائات تکمیل می‌شود.

۴-۱-نظریه استثنائات

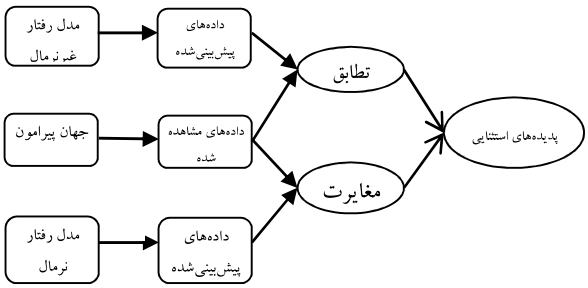
رویکردهای متفاوتی به مسئله استثنائات در حیطه‌های متفاوت علمی و عملی وجود دارد که از آن‌ها می‌توان به استثنائات موضوعی، آماری، ژنتیکی، بیولوژیک و رویکردهای تئوریک اشاره کرد. نخستین‌بار هافمن(۱۹۷۷) تئوری استثنائات را در علم ژنتیک مطرح کرد. پس از ظهور تئوری استثنائات در علم ژنتیک مک کارتی در سال ۱۹۸۰ مفهوم تئوری استثنائات را در استنتاج مطرح کرد. مک کارتی با به‌کارگیری قوانین علی و معلولی به شناسایی استثنائات پرداخت. پس از رویکرد مک کارتی، رویکردهای تئوریک به مسئله شناخت استثنائات مطرح شد. رویکردهای تئوریک به استثنائات براساس یک تئوری که توسط شخصی ایجاد و یا توسعه داده شده، آغاز می‌شود. اگر حیطه نرمال را بتوان برای مسئله تعریف کرد آنگاه استثنائات به‌عنوان شکست در توسعه این تئوری در نظر گرفته می‌شود[23]،[20].

تئوری مفهومی که برای یافتن استثنائات به‌کار گرفته می‌شود، بسته به نوع دانش موجود عمل می‌کند. نقطه شروع مناسب برای توضیح استثنائات در مدل مفهومی انواع مختلف دانشی است که در کاربردهای مختلف استثنائات نقش دارد. دانش ضمنی موجود در سامانه یافتن استثنائات ممکن است بر پایه توضیحی از ساختار نرمال و رفتار وظیفه‌ای سامانه و یا بیان رفتار غیر نرمال سامانه باشد. یافته‌های جمع‌آوری‌شده‌ای که با رفتار نرمال سامانه مطابقت دارند "یافته نرمال " و در غیر اینصورت "یافته غیر نرمال " گفته می‌شود. بر اساس انواع دانش موجود و یافته‌های مشاهده‌شده، دو دسته تئوری شکل گرفته است: تئوری انحراف از ساختار و رفتار نرمال و تئوری انطباق با رفتار غیر نرمال. تئوری انحراف از ساختار و رفتار نرمال تئوری نخستین‌بار توسط ری ریترز (۱۹۸۷) به‌عنوان چارچوب منطقی برای کشف عارضه‌های سامانه‌های فیزیکی با استفاده از مدل ساختار و رفتار نرمال سامانه مطرح شد. براساس این تئوری، استثنائات براساس مقایسه داده‌های مشاهده شده با ساختار و رفتار نرمال سامانه و مغایرت با آن کشف می‌شوند. کشف استثنائات براساس تطابق با رفتار غیرنرمال با درنظرگرفتن دانش رفتارهای استثنایی سامانه عمل می‌کند، به‌نحوی‌که به شبیه‌سازی رفتار غیر نرمال سامانه می‌پردازد.

به‌منظور ارتقای دقت در کشف استثنائات چارچوب تئوریک تلفیقی جدیدی بر اساس تئوری کشف استثنائات براساس سازگاری و تئوری کشف استثنائات بر اساس تطابق با رفتار غیر نرمال به‌صورت زیر (شکل ۱) پیشنهاد می‌شود. استثنائات بر مبنای میزان سازگاری داده‌های مشاهده‌شده با مدل رفتار غیر نرمال و یا مغایرت داده‌های مشاهده‌شده با رفتار نرمال سامانه کشف می‌شوند.

۴-۲-دانه‌بندی اطلاعات

به‌منظور کاهش هدفمند فضای مسأله و تخصیص نسبت بالاتری به داده‌های استثنایی در فضای مورد مطالعه، از تئوری دانه‌بندی اطلاعات استفاده می‌شود.



(شکل-۱): مدل پیشنهادی برای شناسایی داده‌های استثنایی بر اساس تئوری استثنائات

(Figure-1): the proposed model for exceptional data detection on the basis of abnormality theory

دانه‌بندی اطلاعات به‌عنوان ابزاری برای محاسبات تخمینی[1] توسط لطفی‌زاد[2] پیشنهاد شد. دانه‌بندی اطلاعات یک مجموعه از داده، افراز مجموعه مرجع به n دانه اطلاعاتی به‌گونه‌ای است که، داده‌های موجود در یک دانه اطلاعاتی بیشترین شباهت را به یکدیگر داشته باشند. بدین ترتیب یک مسأله دسته‌بندی پیچیده، به چند مسأله ساده‌تر و کوچک‌تر تبدیل شده که منجر به کاهش پیچیدگی فرایند یادگیری می‌شود.

در این رویکرد، فضای مسأله بر اساس میزان شباهت اطلاعات موجود به فضاهای کوچکتر افراز می‌شود. نمونه‌سازی کاهشی با شکستن فضای مسأله، عملکرد مدل طبقه‌بندی‌کننده را بهبود می‌بخشد. هدف از به‌کارگیری تئوری دانه‌بندی اطلاعات، افزایش احتمال کشف داده‌های استثنایی و کاهش پیچیدگی مسأله است.

ساختارهای متعددی برای دانه‌بندی اطلاعات موجود در پایگاه داده وجود دارد؛ اما تمام آن‌ها برای تخمین محتوای انواع داده‌ها کارایی مطلوب را ندارند. انتخاب روش دانه‌بندی مناسب در پروسهٔ کشف استثنائات حائز اهمیت است. یکی از رویکردهای مورد استفاده در دانه‌بندی اطلاعات شناسایی و استخراج دانش موجود در داده در سطوح متفاوت بر اساس میزان شباهت داده‌ها است. بدین منظور از روش خوشه‌بندی K-means استفاده شده است. این نوع خوشه‌بندی به‌دلیل استراتژی‌ای که در جداسازی داده‌ها اتخاذ می‌کند، انتخاب شده است و سعی دارد فضای مسئله را به‌گونه‌ای دسته‌بندی کند که داده‌های هر دسته کم‌ترین فاصله و بیشترین شباهت را به یکدیگر داشته باشند، درحالی‌که دسته‌های مختلف بیشترین فاصله را از یکدیگر داشته باشند. به‌دلیل قرارگیری داده‌های مشابه در یک دسته، دقت قوانین استخراج‌شده از هر دسته افزایش می‌یابد. استثنائات با ساخت دانه‌های اطلاعاتی از داده‌های موجود در پایگاه داده و به‌کارگیری تابع آنتروپی برای هر دانه اطلاعاتی کشف می‌شوند.

۴-۳-آنتروپی

تحلیل محتوا، فنی است که در جستجوی دریافت ادراکات ضمنی قابل استخراج از داده‌هاست. بر این اساس تلاش می‌شود، تا با بررسی یک نقطه داده اطلاعات لازم جهت تحلیل و بررسی رفتار آن اخذ شد[19]،[24]. از اهداف پژوهش حاضر، توسعهٔ استفاده از تئوری اطلاعات به حوزه تحلیل محتوای داده‌ها، برای کشف استثنائات است. نظریّهٔ اطّلاعات مدلی ریاضی از شرایط و عوامل مؤثر در انتقال و پردازش داده ها و اطّلاعات با هدف کمّی‌سازی و اندازه‌گیری عددی اطلاعات است. نخستین‌بار مفهوم آنتروپی توسط *کلاود شانون* (۱۹۴۸) در علم فیزیک و برای کمّی‌سازی مفاهیم عدم قطعیت و بی‌نظمی مطرح شد[22]. این شاخص میزان عدم خلوص مجموعه‌ای از داده‌ها را مشخص می‌کند و با افزایش میزان خلوص داده کاهش می‌یابد. به این معنی که اگر تمام داده‌های موجود در یک پایگاه داده متعلق به یک دسته باشند، میزان آنتروپی برابر با صفر خواهد بود [9]. در این پژوهش، از مفهوم آنتروپی به‌عنوان ابزاری کارآمد برای سنجش میزان استثنایی بودن داده‌ها استفاده شده است.

اگر X یک متغیر تصادفی باشد که یکی از مقادیر x_n و... و x_2 و x_1 را با احتمالات p_n و...و p_2 و p_1 انتخاب می‌کند، تابع $H(x)$ نشان‌دهنده میزان آنتروپی حاصل از آن است که میزان ناخالصی برای بروز حالت مشخص X را نشان می‌دهد. درصورتی‌که پیشامد x_i اتفاق بیفتد، با استفاده از $H(x)$ می‌توان به‌میزان کاهش عدم یقین نسبت به نتایج ممکن دست یافت.

$$H(x) = -\sum_{i=1}^{n} p_i \log(p_i) \qquad (۱)$$

از آنجایی‌که یکی از اهداف این پژوهش، شناسایی داده‌های استثنایی موجود در پایگاه داده است، از تابع آنتروپی رنی به‌عنوان سنجه‌ای برای نمایش میزان اطلاعات و تمایز داده‌های استثنایی از عادی استفاده می‌شود؛ زیرا آنتروپی رنی برای داده‌ها با احتمال وقوع کمتر وزن بیشتری را لحاظ می‌کند و تمایز بین داده‌های عادی و غیرعادی واضح‌تر نشان می‌دهد.

$$H(x) = \frac{1}{1-\propto} \log(\sum p_i^{\propto}) \qquad (۲)$$

[1] Approximate Computing

[2] Lotfi Zade

از تابع آنتروپی رنی با $\propto = 2$ برای یافتن استثنائات استفاده می‌شود. هرچقدر، $\propto$ کوچکتر باشد، توانایی تشخیص داده‌های استثنایی بیشتر است. بدین ترتیب که استفاده از توان ۲ برای احتمالات کوچک باعث بزرگ‌ترشدن مقدار تابع $H(x)$ و وزن‌دهی بیشتر به آن می‌شود.

$$H(x) = -\log(\sum p_i^2) \qquad (۳)$$

آنتروپی یک رکورد نشان‌دهنده میزان نفع اطلاعات موجود در آن سطر از داده است. شود. اگر دو واقعه (X, Y) وابسته باشد آنگاه:

$$H(X, Y) = H(X) + H(Y|X) \qquad (۴)$$

از آن جایی‌که $H(y)$ به‌ازای مقادیر کوچک y بزرگتر است آنتروپی داده‌های غیرنرمال دارای انحراف زیادی نسبت به متوسط تابع آنتروپی دیگر داده‌ها است.

استثنائات با ساخت دانه‌های اطلاعاتی از داده‌های موجود در پایگاه داده و به‌کارگیری تابع آنتروپی برای هر دانه اطلاعاتی کشف می‌شوند. یعنی ابتدا دانه‌های اطلاعاتی توسط روش خوشه‌بندی k-means شکل داده شده و سپس تابع آنتروپی رکوردهای موجود در هر دسته محاسبه می‌شود. رکوردهایی که تابع آنتروپی آن‌ها تفاوت معناداری از آنتروپی سایر رکوردهای موجود در آن دسته دارند، استثنایی خوانده می‌شوند. درنهایت استثنائات موجود در دانه‌های اطلاعاتی تجمیع می‌شوند.

پس از کشف پدیده‌های استثنایی پنهان در پایگاه داده بایستی قوانین بروز رفتار استثنایی در یک سامانه شناسایی شوند. به این منظور از روش پیشنهادی بر اساس رویکرد یادگیری از پایین به بالا استفاده شده است. در ادامه به تشریح روش پیشنهادی در کشف دانش رفتاری استثنائات پرداخته شده است:

۴-۴-استخراج قوانین

اغلب یادگیری رفتار استثنائات فرایند پیچیده‌ای است؛ زیرا رخدادهای استثنایی بسیار کمتر از پدیده‌های معمول اتفاق می‌افتند. وجود حدود مبهم برای بروز رفتار عادی و استثنایی، عدم وجود سنجه‌های ارزیابی مناسب و عدم وجود اطلاعات کافی، از دیگر چالش‌های موجود در فرایند یادگیری رفتار استثنائات است. به منظور غلبه بر کاستی‌های مدل‌های یادگیری معمول، از رویکرد یادگیری از پایین به بالا در قالب روش پیشنهادی RISE ارتقایافته، برای کشف دانش رفتار پدیده‌های استثنایی استفاده می‌شود.

رویکرد یادگیری پایین به بالا نخستین‌بار توسط پدرو دومینگوز در سال ۱۹۹۸ و در قالب الگوریتم یادگیری RISE ارائه شد. در این روش یادگیری قوانین رفتاری با به‌کارگیری تمام داده‌های آموزشی صورت می‌پذیرد؛ به‌طوری‌که ابتدا مطابق با هر رکورد موجود در پایگاه داده یک ضابطه ساخته شده و سپس با در نظرگیری ضوابط مجاور، تعمیم قوانین صورت می‌پذیرد. این رویکرد تمایل زیادی به کشف قوانین خاص و با دقّت بالا دارد؛ اما از آن‌جایی‌که هر رکورد موجود در پایگاه داده به‌عنوان یک ضابطه در نظر گرفته می‌شوند، زمان تعمیم قوانین زیاد است. دومینگوز روش RISE را به همراه سایر روش‌های موجود یادگیری ماشین مانند الگوریتم‌های درخت تصمیم روی سی پایگاه داده استاندارد به‌کار گرفت. این روش یادگیری در مقایسه با سایر روش‌های یادگیری در ۶۶ درصد از موارد عملکرد مطلوب‌تری داشته است. همچنین در سایر موارد از عملکرد قابل قبولی برخوردار بوده است.

بر اساس نوع استراتژی یادگیری، استخراج قوانین به‌طورمستقیم از داده‌های آموزشی صورت گرفته و بدین ترتیب، این استراتژی در شناسایی قوانین کورکورانه عمل می‌کند. در مواجهه با پدیده‌های استثنایی رویکردهای یادگیری جزء به کل بهتر عمل می‌کنند. به همین منظور روش E-RISE تدوین شده است تا ضوابط بروز رفتار استثنایی به‌خوبی فراگرفته شود.

براساس روش پیشنهادی، پس از شناسایی پدیده‌های استثنایی و تفکیک آن از داده‌های عادی، با درنظرگرفتن چارچوب پیشنهادی تئوری استثنائات، با استخراج قوانین حاکم برهر دو دسته دادهٔ عادی و استثنایی آغاز می‌شود. ابتدا داده‌ها به‌تفکیک برچسب عادی و استثنایی دسته‌بندی می‌شود. به‌منظور کاهش هدف‌مند فضای مسأله و افزایش دقّت کشف ضوابط رفتاری، از الگوریتم‌های خوشه‌بندی برای خوشه‌بندی پدیده‌های عادی استفاده می‌شود تا داده‌های نسبتاً مشابه در یک خوشه قرار گیرد. داده‌های استثنایی به‌دلیل تعداد اندک، نیازی به خوشه‌بندی ندارند. پس از آماده‌سازی خوشه‌های عادی و دستهٔ استثنایی، باید ضوابط رفتاری عادی و استثنایی استخراج شود. بدین منظور تعدادی از داده‌های موجود در هر دسته به تصادف انتخاب شده و مطابق با هر داده انتخابی، یک قانون ساخته می‌شود. هر ضابطه شامل n شرط روی n ویژگی تشکیل‌دهنده یک نقطه

داده است. یعنی به‌ازای هر متغیر، یک جمله شرط وجود دارد. هر زوج مرتب(X, Y) معادل یک بردار از صفات به‌صورت $\{x_1, x_2, \dots, x_n\}$ است که نشان‌دهندهٔ ویژگی‌های آن پدیده و متغیر$\{y_j, j = 0,1\}$ برای نشان‌دادن عادی یا نرمال‌بودن آن است. پدیده $\{a_1, a_2, \dots, a_n; 1\}$ را در نظر بگیرید که a_1 تا a_n مقادیر عددی متغیرهای x_1 تا x_n، و عدد یک نشان‌دهنده دسته‌ای است که پدیده مورد بررسی به آن متعلق است. ضابطه اولیه معادل با این پدیده به‌صورت زیر ایجاد می‌شود:

$$\text{IF } x_1 = a_1, x_1 = a_1, \dots, x_1 = a_1 \text{ THEN } y = 1 \qquad (۵)$$

نقاط داده انتخابی به‌عنوان خاص‌ترین قاعده با وجود شرط روی تمامی متغیرها ایجاد شده است؛ درحالی‌که جواب شرط بروز رفتار عادی یا استثنایی را نشان می‌دهد.

پس از شکل‌دهی ضوابط اولیه، باید قوانین به‌گونه‌ای تعمیم یابند که بتواند رفتار پدیده‌های موجود در آن پایگاه داده را توجیه کنند. بنابراین، به‌منظور افزایش دقّت پیش‌بینی با استفاده از شاخص فاصله، ضوابط اولیه به‌گونه‌ای تعمیم داده می‌شود تا نقاط داده‌ای که در همسایگی ضابطه انتخابی قرار دارند، پوشش داده شوند. بدین ترتیب که ضوابط، یک‌به‌یک مورد بررسی قرار می‌گیرد و عمومی‌سازی با استفاده از نقاط نزدیک به ضابطهٔ مورد بررسی موجود در آن دسته که تاکنون توسط ضابطهٔ دیگری پوشش داده نشده‌اند، انجام می‌شود.

هدف از مرحله تعمیم، تعدیل ضابطه ایجادشده با حداقل تغییر به‌صورتی است که دقّت پیش‌بینی در سطح قابل قبولی حفظ شود. در مواجهه با دادهٔ جدید، مقدار دادهٔ مورد بررسی در هر متغیر با حدود رفتاری بالا و پایین مقایسه می‌شود؛ در صورتی که داده مورد بررسی در بازه‌های موجود باشد، تغییری در قوانین صورت نمی‌گیرد؛ اما اگر بازه‌های طراحی‌شده نتواند داده همسایه را پوشش دهد، با بررسی دو سنجه میانگین هندسی دقت مثبت و منفی و میزان پوشش‌دهی، میزان بهبود تعمیم ایجاد شده در ضابطه مورد بررسی ارزیابی می‌شود. درصورتی‌که تغییر ایجادشده مطلوبیت لازم را داشته باشد، اعمال شده و در غیر این صورت تغییرات در نظر گرفته نمی‌شود. برای تعیین شعاع همسایگی از روش نزدیک‌ترین همسایگی در شعاع ثابت استفاده می‌شود.

به‌منظور سنجش کارایی الگوریتم پیشنهادی، از دو شاخص G-means و میزان پوشش‌دهی استفاده می‌شود. این سنجه در پژوهش‌های زیادی برای ارزیابی عملکرد الگوریتم‌های یادگیری در مجموعه داده‌های نامتوازن، استفاده شده است که دقت قوانین پیش‌بینی‌کنندهٔ دسته‌های اصلی و فرعی را مورد بررسی قرار می‌دهد.

$$\text{G} - \text{means} = \sqrt{\frac{d}{c+d}.\frac{a}{a+b}} \qquad (۶)$$

با استفاده از معیار ارزیابی G-means و درجه پوشش‌دهی، تغییرات درصورتی لحاظ می‌شوند که شاخص G-means بدتر نشده و قانون ایجادشده درصد مشخصی از داده‌های موجود در آن دسته را بپوشاند.

از مزیت‌های روش پیشنهادی این است که مرحله تعمیم در هر بار برای یک متغیر خاص انجام نمی‌شود؛ بلکه مانند روش رویکرد خوشه‌بندی، تمام متغیرها را در نظر می‌گیرد. به‌کارگیری استراتژی پوشش متوالی، احتمال بروز مشکل تجزیه اطلاعات را کاهش می‌دهد. نتیجه این اقدامات، ضوابط رفتاری پدیده‌های عادی و استثنایی به‌صورت یک پایگاه دانش است. مزیت روش پیشنهادی یادگیری، استفاده از تمام نقاط داده موجود در پایگاه داده و ساخت موازی ضوابط است؛ درحالی‌که در رویکرد یادگیری از بالا به پایین در هر زمان با درنظرگرفتن یک متغیر فقط یک ضابطه ساخته می‌شود که منجر به ازدست‌رفتن اطلاعات می‌شود. روش پیشنهادی با کمینه‌کردن مشکل تجزیه اطلاعات مانع ازدست‌رفتن آن می‌شود. فرایند استخراج دانش رفتاری پدیده‌های عادی و استثنایی در دو فاز ایجاد قوانین و تعمیم در جدول (۱) تلخیص شده است.

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

گام ۱-افراز فضای مسأله برمبنای میزان شباهت اطلاعاتی داده‌ها با استفاده از روش خوشه‌بندی K-means

گام ۲- محاسبه آنتروپی داده‌های موجود در هر خوشه

گام ۳- شناسایی داده‌های استثنایی هر خوشه- داده‌هایی که آنتروپی آن‌ها تفاوت معناداری از میانگین آنتروپی داده‌های آن خوشه دارد.

گام ۴- تجمیع استثنائات شناخته‌شده هر خوشه

گام ۵- دسته‌بندی داده‌ها- ابتدا داده‌های نرمال و استثنایی شناسایی‌شده در گام دو را در دو دسته جداگانه قرار داده و سپس داده‌های نرمال دوباره به‌گونه‌ای

خوشه‌بندی می‌شود تا داده‌های نسبتاً مشابه در یک خوشه قرار گیرند.

گام ۶-درصدی (α %) از داده‌های موجود در هر خوشه ایجاد شده (گام ۵) را به تصادف انتخاب کرده و به‌ازای هر داده موجود در نمونه تصادفی، یک قانون ایجاد می‌شود.

گام ۷-قوانین ایجاد شده 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 بازکردن بازه‌ها برای متغیرهای عددی انجام می‌گیرد. به‌منظور سنجش کارایی فرایند عمومی‌سازی در هر مرحله از شاخص G-means و support استفاده می‌شود. تغییرات درصورتی لحاظ می‌شوند که شاخص G-means بدتر نشده و قانون ایجادشده درصد مشخصی از داده های موجود در آن دسته را بپوشاند.

گام ۸-طراحی سیستم خبره برای شناسایی استثنائات جدید با استفاده از چارچوب پیشنهادی تئوری استثنائات. استثنائات جدید براساس قوانین رفتاری نرمال و استثنایی کشف می‌شود. بدین ترتیب، داده‌هایی که با قوانین رفتار نرمال مطابقت ندارند یا منطبق با قوانین رفتار استثنایی عمل می‌کنند، به‌عنوان پدیده‌های استثنایی تلقی می‌شوند. شکل (۲) فرایند کشف و یادگیری از رفتار استثنائات را نشان می‌دهد.

۵-یافته‌های پژوهش

در پژوهش حاضر، کشف پدیده‌های استثنایی و یادگیری رفتار آن‌ها در قالب مدل پیشنهادی و به‌منظور افزایش اطمینان به پدیده‌های استثنایی، انجام می‌پذیرد. کارایی الگوی پیشنهادی با مطالعهٔ رفتار سهام موجود در بازار اوراق بهادار ایران با هدف شناسایی سهام استثنایی و یادگیری رفتار آن‌ها، بررسی شده است. بازار سهام به‌دلیل داشتن پتانسیل بالای سودآوری همواره مورد توجه سرمایه‌گذاران بوده است. این بازار دارای سامانه‌ای پیچیده و پویا با داده‌هایی ناهمگون و بی‌نظم است.

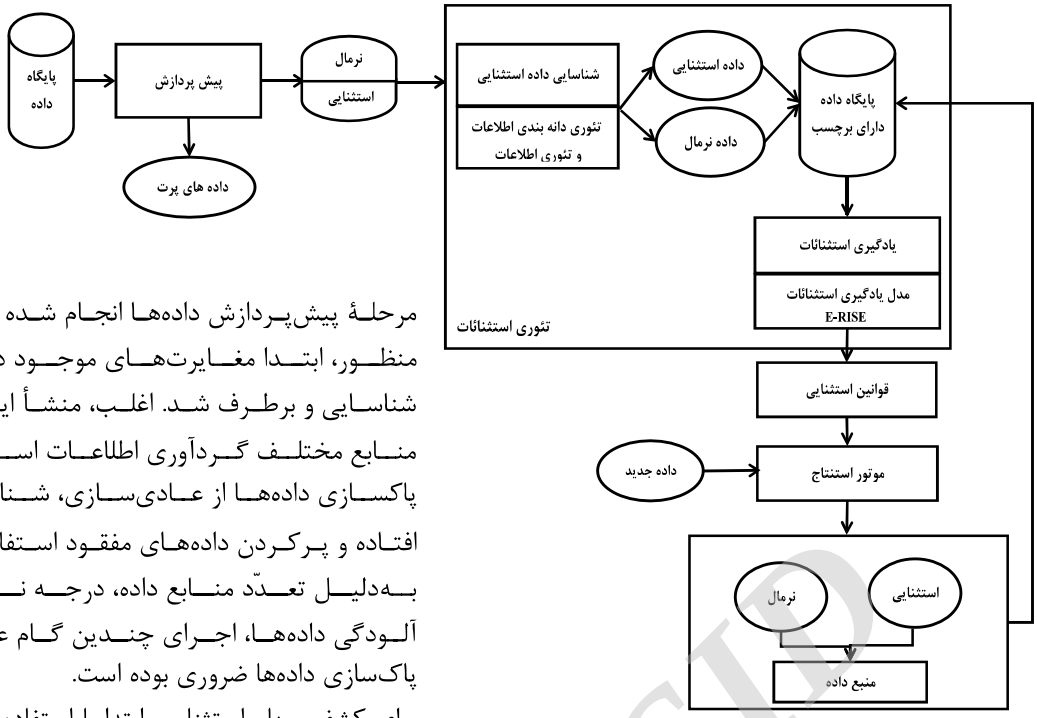
وجود ابزارهای گوناگون شناسایی و انتخاب سهام، روند سرمایه‌گذاری را با چالش جدّی مواجه ساخته است؛ زیرا به‌کارگیری عوامل مختلف باعث ایجاد نتایج متفاوت می‌شود. از سوی دیگر، مسأله عدم شفافیت و قطعیت در بازار سهام، انتخاب سبد سهام استثنایی را به یک معضل اساسی در حوزه تصمیم‌گیری مالی، مبدّل کرده است. با این تفاسیر، یافتن مجموعهٔ استثنایی از سهام که منفعت استثنایی برای سهام‌دار در پی داشته باشد، یکی از دغدغه‌های مهم فعالان بازار سرمایه محسوب می‌شود. سهام استثنایی رفتاری متفاوت با الگوهای اصلی و مورد انتظار سهام موجود در بازار اوراق بهادار، از خود بروز می‌دهند و اغلب، شناسایی آن‌ها امری دشوار است؛ زیرا سهام غیرعادی حجم کمی از سهام موجود در بازار سهام را شامل می‌شود و روش‌های معمول داده‌کاوی توانایی شناسایی الگوهای رفتاری آن‌ها را ندارند.

(جدول-۱): مراحل استخراج قوانین بر اساس الگوریتم E-RISE

(Table-1): Rule extraction based on E-RISE algorithm

مرحلهٔ عمومی‌سازی	مرحلهٔ ایجاد قوانین
اگر $R =$ یک $(a_1, a_2, \dots, a_m, C_R)$ رول و $N =$ یک $(e_1, e_2, \dots, e_m, C_n)$ 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

با هدف تشخیص سهام غیرعادی بر اساس مطالعات پیشین و نظر خبرگان متغیرهای موثر در بروز رفتار استثنایی سهم، بازده، نسبت قیمت به سود، حجم معاملات، دفعات معاملات، ارزش معاملات، کمترین قیمت و بیشترین قیمت شناسایی شده‌اند. اطلاعات مالی در محدوده این هفت ویژگی برای شرکت‌های فعال در بازار اوراق بهادار ایران به‌صورت ماهیانه، از نیمه دوم سال ۱۳۹۰ تا پایان سال ۱۳۹۲در ۶۰۲۲ رکورد، گردآوری شده است. به‌منظور تشکیل پروفایل از سهام مورد بررسی و کشف استثنائات، لازم است مجموعهٔ قابل نمایشی از سهام ایجاد شود.



(شکل-۲):روش پیشنهادی بر پایه رویکرد تلفیقی تئوری استثنائات و اطلاعات

(Figure-2): the proposed learning methodology based on abnormality and information theories

فرض کنید $x_{i,j}$ نشان‌دهنده سهم i ام در مشخصه j باشد، درحالی‌که $i = \{۱و۲و...و۶۰۲۲\}$ نشان‌دهنده ۶۰۲۲ سهم مورد بررسی و $j = \{۱و۲و...و۷\}$ متغیرهای درنظرگرفته‌شده در شناسایی و یادگیری رفتار استثنایی سهام است. در این‌صورت M نشان‌گر پروفایل سهام مورد بررسی خواهد بود.

$$M = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,7} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,7} \\ \vdots & \vdots & \cdots & \vdots \\ x_{6022,1} & x_{6022,2} & \cdots & x_{6022,7} \end{bmatrix} \quad (۱-۴)$$

به‌دلیل اهمیت دقیق‌بودن ورودی‌های مدل و تأثیر آلودگی داده‌ها بر تنزّل عملکرد فرایند داده‌کاوی، مرحلهٔ پیش‌پردازش داده‌ها انجام شده است. به این منظور، ابتدا مغایرت‌های موجود در پایگاه داده، شناسایی و برطرف شد. اغلب، منشأ این مغایرت‌ها منابع مختلف گردآوری اطلاعات است. به‌منظور پاکسازی داده‌ها از عادی‌سازی، شناخت نقاط دور افتاده و پرکردن داده‌های مفقود استفاده شده است. به‌دلیل تعدّد منابع داده، درجه ناهمگن‌بودن و آلودگی داده‌ها، اجرای چندین گام عملِ تبدیل و پاک‌سازی داده‌ها ضروری بوده است.

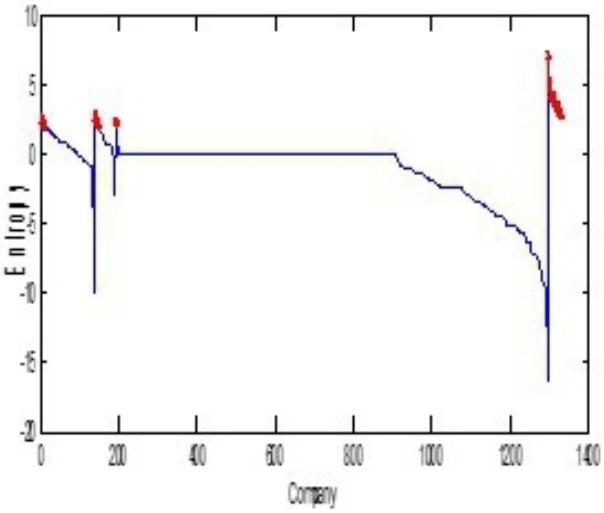
برای کشف سهام استثنایی، ابتدا با استفاده از "نمونه‌سازی کاهشی"، دانه‌های اطلاعاتی ساخته می‌شود. روش خوشه‌بندی K-means برای تشکیل دانه‌های اطلاعاتی به‌کارگرفته شده و درنتیجه، سه خوشه ایجاد شده است. بدین ترتیب، فضای مسأله برمبنای میزان شباهت اطلاعاتی داده‌ها، به فضاهای کوچک‌تر، افراز می‌شود که باعث بهبود عملکرد مدل طبقه‌بندی خواهد شد؛ سپس، از تابع آنتروپی رنی برای تشخیص بی‌نظمی‌های هر دانه اطلاعاتی استفاده می‌شود. براساس تابع آنتروپی، فضای هر دانهٔ اطلاعاتی به دو دستهٔ عادی و استثنایی تقسیم شده است.

در پژوهش حاضر، سهام استثنایی سهامی هستند که میزان تابع آنتروپی آنها به فاصله $+۱/۵\sigma$ از میانگین تابع آنتروپی تمامی سهام است. شکل (۳) تابع آنتروپی سهام ترسیم شده است. همان‌گونه‌که در تصویر مشخص است، میزان آنتروپی سهام استثنایی با میانگین آنتروپی سایر سهام تفاوت معناداری دارد.

(جدول-۲): مشخصه سهام مورد بررسی

(Table-2): stocks characteristics

مشخصه	بازده (درصد)	قیمت/سود	حجم معاملات (تعداد سهام)	دفعات معاملات	ارزش معاملات (ریال)	کمترین قیمت (ریال)	بیشترین قیمت (ریال)
میانگین	0.297	16.8	5157980.1	173.2	4.0E+10	6103.1	6273.3
انحراف معیار	1.131	80.6	181004543	565	1.78E+12	7739.5	7963.3
حداقل	-5.317	-109	0	0	0	0	0
حداکثر	44.59	2842	9.9E+9	22772	9.80E+13	80000	83899



(شکل-۳): تابع آنتروپی سهام
(Figure-3): stock entropy function

چنان‌که پیش از این نیز اشاره شد، به‌منظور استخراج قوانین حاکم بر بروز رفتار استثنایی داده‌ها، از رویکرد یادگیری از پایین به بالا استفاده شده است. بر اساس این روش، ابتدا داده‌های استثنایی موجود در خوشه‌ها جدا می‌شوند و در یک خوشه مستثنی قرار می‌گیرند. بدین ترتیب، سهام موردنظر در چهار گروه خوشه‌بندی شده است که شامل یک خوشه از داده‌های استثنایی و سه خوشه از داده‌های عادی می‌شود. استفاده از خوشه‌بندی باعث می‌شود داده‌های نسبتاً مشابه در یک گروه قرار گیرند که باعث بالارفتن دقّت قوانین مستخرج از آن خوشه می‌شود. اطلاعات مربوط به خوشه‌ها در جدول (۳) خلاصه شده است.

درصدی (α %) از سهام موجود در هر خوشه به‌طور تصادفی، انتخاب و به‌ازای هر داده، یک ضابطه منطبق بر آن ساخته شده است. قابل ذکر است که مقایر متفاوتی برای α در خوشه‌های مختلف، لحاظ شده است. جدول (۴) تاثیر تغییر در مقدار α را بر دقّت پیش‌بینی و تعداد قانون استخراج‌شده نشان می‌دهد. با توجه به اطلاعات جدول (۳) بهترین مقدار برای α در خوشه‌های نخست و دوم و سوم و چهارم به‌ترتیب برابر با ۳۰٪، ۱/۵٪، ۰/۴٪ و ۰/۵٪ است؛ سپس براساس رویکرد یادگیری از پایین به بالا و با استفاده از تابع فاصله اقلیدسی، ضوابط اولیه تعمیم می‌یابد تا تمام داده‌ها پوشش داده شود. ضوابط جدید با به‌کارگیری داده‌های مجاور ضابطه اولیه، که تا به‌حال توسط هیچ یک از قوانین موجود پوشش داده نشده‌اند، بسط داده می‌شود. این کار با درنظر گرفتن بیشینه دقّت و کمینه تعمیم، از طریق حذف شرایط یا گسترش بازه‌ها برای متغیرهای عددی، صورت می‌پذیرد. درصورتی‌که اندازه فاصله کم باشد، تعداد تکرارها و زمان تعمیم افزایش می‌یابد و درصورتی‌که فاصله زیاد باشد، دقّت قوانین کاهش خواهد یافت. با درنظرگرفتن فاصله همسایگی مناسب، الگوی رفتاری استثنائات در قالب سه ضابطه موجود در جدول (۵) شناسایی شده است.

(جدول-۳): اطلاعات خوشه‌ها
(Table-3): clusters information

مشخصه	بازده	قیمت/ سود	حجم معاملات	دفعات معاملات	کمترین	بیشترین	ارزش معاملات
خوشه ۱- استثنایی (۹)							
میانگین	2.60	21.84	2.7E+8	222.7	25047	25726	2/76E+12
انحراف استاندارد	5.71	64.14	1.64E+9	491.5	19005	19541	1/62E+13
خوشه 2-عادی (۵۲۳)							
میانگین	0.21	15.15	840737	112	3962	4068	2/7E+9
انحراف استاندارد	0.79	1.34	1759264	184	3090	3168	6/2E+9
خوشه ۳-عادی(۴۶۲۸)							
میانگین	0.483	7.63	354430	145	22686	23350	8/5E+9
انحراف استاندارد	0.89	5.37	819099	268	9471	9772	2/2E+10
خوشه ۴-عادی (۸۲۷)							
میانگین	0.89	-9.71	2.1E+7	1520	3979	4115	8/5E+10
انحراف استاندارد	2.04	20.34	2.8E+7	2561	4132	4258	1/3E+11

(جدول-۴): تعیین تعداد بهینه قوانین مستخرج

(Table-4): optimal number of extracted rules

خوشه	تعداد داده	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8
۱	۹	0.5	0.4	0.4	0.4	0.3	0.3	0.3	0.4
۲	۴۶۲۸	0.008	0.007	0.006	0.005	0.004	0.003	0.002	0.0015
۳	۸۲۷	0.01	0.01	0.008	0.006	0.006	0.005	0.005	0.004
۴	۵۲۳	0.03	0.03	0.02	0.02	0.01	0.01	0.005	0.005
تعداد قانون استخراجی		65	59	48	42	31	25	18	16
G-means (%)		52.35	53.46	60.37	59.87	64.87	70.09	83.74	86.53

(جدول-۵): قوانین بروز رفتار عادی سهام استخراج‌شده توسط الگوریتم RISE ارتقایافته

(Table-5): exceptional stocks behavioral rules extracted by E-RISE algorithm

اگر	آنگاه سهام
قیمت به سود> ۶/۴۶ و بازده>۴۴/۵۹ و حجم معاملات>۵۵۶و ارزش معاملات >۹۴۶۵۶۰۰ و قیمت>۱۷۰۲۵	استثنایی
بازده>۰/۶۴ و ارزش معاملات > ۹۸۰۲۴۰۹۴۹۶۴۳۳۶و حجم معاملات > ۹۹۱۵۴۴۵۵۷۶	استثنایی
بازده> ۰/۲۶ و ارزش معاملات> ۲۷۴۸۷۹۵۶۲ و قیمت> ۴۹۲۰۸	استثنایی

با استخراج قوانین حاکم بر رفتار عادی و استثنایی سهام، به‌منظور پیش‌بینی وضعیت سهام جدید، سامانه خبره شناسایی سهام استثنایی در محیط نرم‌افزار GURU طراحی شده است. این سامانه با استفاده از قوانین شناسایی‌شدهٔ حاکم بر رفتار استثنایی سهام به شناسایی سهام استثنایی جدید می‌پردازد. در مواجهه با داده جدید سامانه خبره فعالیت خود را آغاز کرده و قوانین مورد آزمایش قرار گرفته تا نوع داده جدید تشخیص داده شود. این سامانه توانایی آزمودن یک سهم خاص و یا جستجو در یک پایگاه داده و شناسایی سهام استثنایی را دارد. سامانه خبرهٔ طراحی‌شده به‌منظور کشف سهام استثنایی موجود در پایگاه دادهٔ سهام، به‌کار گرفته شده است تا قابلیت قواعد استخراج‌شده توسط الگوریتم RISE ارتقایافته، سنجیده شود. نتایج حاصل از به‌کارگیری سامانه خبره طراحی شده روی پایگاه دادهٔ سهام، نشان‌دهندهٔ قابلیت روش پیشنهادی در کشف سهام استثنایی و یادگیری رفتار آن‌هاست.

به‌منظور ارزیابی عملکرد مدل پیشنهادی کشف و یادگیری از استثنائات، عملکرد الگوریتم‌های یادگیری C4.5 ، CART و CHAID روی پایگاه داده پاک‌سازی شده با استفاده از نرم‌افزار کلمنتاین ۱۲ مورد بررسی قرار گرفته و مدل پیش‌بینی رفتار عادی و استثنایی ساخته شده است. جدول (۶) میزان تابع G-means و دقت را برای این الگوریتم‌ها نشان می‌دهد.

الگوریتم دسته‌بندی CHAID تمامی داده‌ها را به‌عنوان دادهٔ عادی شناسایی و قوانین رفتاری آن‌ها را کشف کرده است. الگوریتم CART تنها ۳۵ پدیدهٔ استثنایی را به‌درستی پیش‌بینی کرد درحالی‌که از لحاظ دقت پیش‌بینی ۹۷/۵۶ درصد از پدیده‌ها را به‌درستی شناسایی کرده است. این الگوریتم یک درخت در دو سطح ایجاد کرده و از متغیر تعداد دفعات معاملات برای تفکیک الگوی رفتاری پدیده‌های عادی و استثنایی استفاده کرده است.

الگوریتم C4.5 از بین یکصد و پنجاه سهم استثنایی، پنجاه و سه سهم استثنایی را به‌درستی شناسایی کرده است؛ اما برای ایجاد قوانین فقط از سه متغیر نسبت قیمت به سود، دفعات معاملات و قیمت استفاده کرده است.

خروجی‌های حاصل از مدل‌های به‌کار گرفته‌شده، نشان می‌دهد این الگوریتم‌ها برخلاف دقت بالا در پیش‌بینی، به‌دلیل تمایلشان به کلیت، فقط قادر به پیش‌بینی الگوی رفتاری دسته‌های اصلی هستند. الگوریتم E-RISE در کنار برآوردن دقت بالا، شاخص G-means بالاتری نسبت به بقیه الگوریتم‌های مورد بررسی داشته است.

۶-نتیجه‌گیری

هدف از این پژوهش، طراحی ابزاری کارامد برای کشف و شناسایی داده‌های استثنایی از میان انبوه داده موجود در پایگاه داده و یادگیری رفتار آن‌ها است. در پژوهش حاضر، رویکرد تدوین مجدد مسأله اصلی برگزیده شده است؛ به‌این

ترتیب که استثنائات شناسایی می‌شود و در یک دستهٔ جداگانه قرار می‌گیرد؛ سپس، با به‌کارگیری الگوریتم E-RISE، قوانین رفتار عادی و استثنایی استخراج می‌شود. این روش یادگیری، به‌دلیل عدم شکست فضای مسأله و به‌کارگیری تمام متغیرها در فرایند یادگیری، مانع از دست رفتن اطلاعات می‌شود. همچنین فرایند عمومی‌سازی در هر بار برای یک متغیرخاص، انجام نمی‌شود؛ بلکه این روش، تمام متغیرها را در نظر می‌گیرد؛ بدین صورت که انتخاب دادهٔ نـــزدیک از بین تمام نقاط موجود در یک دسته، انجام می پذیرد.

به‌منظور سنجش کارایی روش پیشنهادی به کشف و یادگیری سهام استثنایی موجود در بازار بورس ایران پرداخته شده است. اطلاعات مربوط به شرکت های حاضر در بازار بورس تهران در بازه زمانی ۱۳۹۰- ۱۳۹۲ در هفت متغیر بازده، نسبت قیمت به سود، حجم معاملات، دفعات معاملات، ارزش معاملات، کمترین قیمت و بیشترین قیمت مورد برسی قرار گرفت. از قوانین استخراجی برای پیش‌بینی رفتار سهام جدید و تشکیل پرتفوی بهینه استفاده می‌شود.

پیشنهاد می‌شود به‌منظور افزایش دقت تشخیص آستانهٔ رفتار استثنایی سهام از منطق فازی در تعیین آستانه‌های قابل قبول بروز رفتار استثنایی به‌کار گرفته شود که می‌تواند به افزایش دقّت فرایند شناسایی استثنائات منجر شود.

(جدول-۶): ماتریس درهم‌ریختگی

(Table-6): confusion matrix

روش	درست		غلط		دقت	G-means
	نرمال	استثنایی	نرمال	استثنایی		
CHAID	5811	0	25	150	97.05	0
CART	5800	35	36	105	97.65	69.75
C4.5	5811	53	25	97	97.95	82.72
E-RISE	5815	137	23	13	99.4	98.72

۷-مراجع 7-Refrences

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

[2] J. Boshes, Change point detection in cyber-attack data, Ph.D. dissertation, Arizona state university, 2009.

[3] J. Burez and D. Van den Poel, "Handling class imbalance in customer churn prediction", *Expert Systems with Applications,* vol. 36, pp. 4626–4636, 2009.

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

[5] L. Cao, Y. Zhao and C. Zhang, "Mining Impact-Targeted Activity Patterns in Imbalanced Data", *IEEE Transactions on knowledge and data engineering*, vol. 20, pp.1053-1066, 2008.

[6] N .V. Chawla, N. Japkowicz and A .K. lcz, "Editorial: Special Issue on Learning from Imbalanced Data Sets", *SIGKDD Explorations*, vol. 6, pp.1–6, 2004.

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

[8] E. Clark, "Exploiting stochastic dominance to generate abnormal stock returns", *Journal of Financial Markets*, vol. 20, pp.20–38, 2014.

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

[10] T .V. Duong, H .H. Bui, D .Q. Phung and S. Venkatesh, "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.

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

Entropy Computation, Research Journal of Applied Sciences", *Engineering and Technology*, vol. 8, pp. 398-409, 2014.

[25] G. Weiss, "Mining with rarity: A unifying framework", *SIGKDD Explorations Special Issue on Learning from Imbalanced Datasets*, vol. 6, pp. 7–19, 2004.

[26] T. Xiang and S. Gong, "Video Behavior Profiling for Anomaly Detection", *IEEE Trans. on Pattern Analysis and Machine* Intelligence, vol 30, pp. 893–908,2008.

الهه حاجی‌گل یزدی دانش آموخته مقطع دکترای دانشگاه یزد در رشته مهندسی صنایع است که در زمینه داده-کاوی، یادگیری ماشین و طراحی سامانه‌های اطلاعاتی فعالیت می‌کند. علاقه‌مندی ایشان، یادگیری رفتار پدیده‌های استثنایی و تحلیل بازارهای مالی است.

نشانی رایانامه ایشان عبارت است از:

elahehajigol@gmail.com

مسعود عابسی دانش‌آموخته دانشگاه ایالتی کلمسن آمریکا در رشته مدیریت صنعتی، و در حال حاضر عضو هیئت علمی دانشکده مهندسی صنایع یزد، مدیر گروه مدیریت دانش و هوشمندسازی مرکز مطالعات دانشگاه تربیت مدرس، و مشاور داده‌کاوی مدیرعامل سازمان بیمه سلامت ایران است. وی قبلاً مشاور داده‌کاوی مدیرعامل بانک اقتصاد نوین، سازمان پشتیبان تأمین اجتماعی، شرکت سوخت‌رسانی هواپیمایی اوج و مشاور مدیرعامل شرکت فولاد آلیاژی ایران بوده است. نامبرده استاد مدعو دانشگاه دولتی مالزی، دانشگاه تهران و دانشگاه اسکرانتون ایالت پنسیلوانیای آمریکا بوده است.

نشانی رایانامه ایشان عبارت است از:

Mabessi@yazd.ac.ir

محمدباقر فخرزاد دانش‌آموخته دانشگاه علم و صنعت ایران در رشته مهندسی صنایع است. در حال حاضر ایشان عضو هیأت علمی و دانشیار دانشکده مهندسی صنایع دانشگاه یزد بوده و در زمینه‌های مدیریت زنجیره تأمین، زمانبندی تولید

[12] R.S. Gong, "A Segmentation and Re-balancing Approach for Classification of Imbalanced Data", Ph.D. dissertation, University of Cincinnati, 2010.

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

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

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

[16] M. V. Joshi, "Learning Classifier Models for Predicting Rare Phenomena", Ph.D. dissertation, University of Minnesota, Twin Cites, Minnesota, USA, 2002.

[17] Y. Kim and S. Y. Sohn, "Stock fraud detection using peer group analysis", *Expert Systems with Applications,* vol. 39, pp. 8986–8992, 2012.

[18] Y. Kou, "Abnormal Pattern Recognition in Spatial Data", Ph.D. dissertation, Faculty of Virginia Polytechnic Institute and State University, 2006.

[19] X. Li and F. Rao, "Outlier Detection Using the Information Entropy of Neighborhood Rough Sets",*Journal of Information & Computational Science*, vol. 9, pp. 3339–3350, 2012.

[20] J. McCarthy, "Applications of circumscription to formalizing common-sense knowledge", *Artificial Intelligence*, vol. 28, pp. 89-116, 1986.

[21] J. Nagi, "An intelligent system for detection of non-technical losses in Tanaga National Berhad (TNB) Malaysia low voltage distribution network", Ph.D. dissertation, Tenaga national university,2009.

[22] U. Qamar, "Automated Entropy Value Frequency (AEVF) Algorithm for Outlier Detection in Categorical Data", *Recent Advances in Knowledge Engineering and Systems Science*, pp. 28-35, 2011.

[23] R. Reiter, "A Theory of Diagnosis from First Principles", *Artificial Intelligence*, vol. 32, pp. 57-95, 1987.

[24] D. B. Setyohadi, A. Abu Bakar and Z. A. Othman, "Rough K-means Outlier Factor Based on

و داده‌کاوی فعالیت می‌کند.
نشانی رایانامه ایشان عبارت است از:
mfakhrzad@yazd.ac.ir

حسن حسینی‌نسب دانش‌آموخته دانشگاه باث انگلستان در رشته مهندسی صنایع است. در حال حاضر ایشان با مرتبه علمی استاد تمام عضو هیأت علمی دانشکده مهندسی صنایع دانشگاه یزد بوده و در زمینه برنامه‌ریزی تولید، طراحی سامانه‌های تولیدی، سامانه‌های ساخت و تولید پیشرفته و برنامه‌ریزی استراتژیک فعالیت می‌کند.
نشانی رایانامه ایشان عبارت است از:
hhn@yazduni.ac.ir