

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

الهه حاجی گل 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 رخ‌می‌دهد 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 به عنوان داده‌های پرت شناسایی کرده و از محاسبات خارج می‌کنند و یا اینکه به‌دلیل تمایل به کلیت، قابلیت شناسایی و دسته‌بندی درست این پدیده‌ها را ندارند. به همین دلیل، ایجاد چارچوبی کارآمد برای کشف دانش و یادگیری رفتار پدیده‌های استثنایی محدود 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 ماهه سهام شرکت‌های فعال در بازار اوراق بهادار ایران به منظور 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 ( $\mathbb{E}$  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.

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 خوشبندی، روش‌های آماری، نزدیکترین همسایگی و ... برای شناسایی استثنایات 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؛ بهنحوی که بتوان مسئله 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], [7], [17].

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

روش‌های حساس به هزینه با درنظرگرفتن هزینه 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 داده‌های 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. استثنایات خود از لحاظ ماهیت رفتاری و

## ۱-مقدمه

منطق یادگیری استثنایات به عنوان 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 تلقی می‌شوند.

در تعریف پدیده‌های استثنایی، دیدگاه‌های مختلفی موجود است؛ نخستین بار دیویس در سال ۱۹۷۷، پدیده‌های نادر را در قالب مسئله عارضه‌بایی سامانه‌های فیزیکی و براساس انحراف از ساختار و رفتار طبیعی، شناسایی کرد. پس از آن، هنری کوهن (۱۹۸۱) "غیرعادی‌بودن" را به عنوان "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 استثنایات شود؛ سپس ریترز و مک‌کارتی شکسته‌های یک سامانه فیزیکی را حالات غیر طبیعی سامانه تلقی کرده و با به کارگیری قوانین علی 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 می‌دهند.

کشف دانش پدیده‌های استثنایی مسئله‌ای چالش‌برانگیز است. یک پایگاه داده بزرگ شامل انبوی از رکوردهاست که اغلب از الگوهای رفتاری مورد انتظار سامانه تبعیت می‌کنند. در این میان تعداد بسیار محدودی پدیده 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 به عنوان داده‌های پرت شناسایی کرده و از محاسبات خارج می‌کند و یا اینکه به دلیل 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

فصلی



در مسأله شناسایی داده‌های استثنایی و یادگیری قوانین رفتاری آن‌ها پیشنهاد شده است؛ به طوری که 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 استثنایات و قراردادن آن‌ها در دسته جدآگانه، رویکرد یادگیری از پایین به بالا و تنویر استثنایات برای یادگیری رفتار استثنایات استفاده شده است؛ به طوری که قابلیت به کارگیری 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 استثنایات، پدیده‌های استثنایی کشف شده و در دسته جدآگانه‌ای قرار می‌گیرد؛ سپس با استفاده از رویکرد یادگیری از پایین به بالا و براساس الگوریتم پیشنهادی RISE ارتقا یافته، رفتار داده‌های عادی و استثنایی، فراگرفته می‌شود. با استفاده از این رویکرد تمام پدیده‌های موجود در پایگاه داده مورد بررسی قرار گرفته و قوانین رفتاری به طور 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 استخراج شده در قالب یک سامانه خبره به‌منظور پیش‌بینی داده‌های استثنایی جدید استفاده می‌شود. به‌کارگیری مدل ارائه شده برای شناسایی استثنایات 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 استخراج شده در قالب چارچوب تلفیقی پیشنهادی تئوری استثنایات، برای طراحی سامانه خبره به‌کار گرفته می‌شود تا چرخه کشف، یادگیری و تبدیل اطلاعات کامل شود.

به کارگیری روش پیشنهادی موجبات تحقق اهداف زیر فراهم می‌آورد:

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

به کارگیری مدل پیشنهادی برای شناسایی سهام استثنایی و یادگیری قوانین رفتاری سهام استثنایی.

اهمیت با هم متفاوتند؛ لذا در نظر گرفتن یک هزینه برای دسته‌بندی اشتباه استثنایات نتیجه دلخواه را ایجاد نمی‌کند. به کارگیری روش‌های جستجوی غیرطعام و به طور خاص الگوریتم ژنتیک، به‌دلیل ترسیم فضای مسأله متناسب با یادگیری رفتار استثنایی سامانه، انتخاب تابع مطلوب‌بیت مناسب و تنظیم پارامترهای بهینه برای شناسایی استثنایات 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 " به عنوان یک تکنیک " دسته‌بندی نیمه‌نظری " 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 فرآگیری رفتار استثنایات 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 دارد. این روش‌ها عملکرد الگوریتم‌های دسته‌بندی سنتی را در یادگیری استثنایات ارتقا بخشنده. تلفیق این روش‌ها و سایر روش‌های یادگیری استثنایات، باعث افزایش احتمال شناسایی داده‌های استثنایی و بهبود دقیق یادگیری می‌شود [21, 7].

چالش‌های دیگری نیز در مسأله 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 استثنایات اشاره کرد [25]. از این‌رو 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 استثنایات از رویکرد یادگیری از بالا به پایین، برای یادگیری استفاده کردند. این شیوه به‌دلیل شکست فضای مسئله براساس اطلاعات کسب شده از متغیرها، باعث ازدست‌رفتن اطلاعات می‌شود. با در نظر گرفتن پژوهش‌های صورت گرفته در زمینه کشف استثنایات و یادگیری رفتار آنها [4], [11], [23], [25] رویکرد نوینی برای حل چالش‌های موجود

پرداخت. پس از رویکرد مک کارتی، رویکردهای تئوریک به مسئله شناخت استثنایات مطرح شد. رویکردهای تئوریک به استثنایات براساس یک تئوری که توسط شخصی ایجاد و یا توسعه داده شده، آغاز می‌شود. اگر حیطه نرمال را بتوان برای مسئله تعریف کرد آنگاه استثنایات به عنوان شکست در توسعه این تئوری در نظر گرفته می‌شود[23],[20].

تئوری مفهومی که برای یافتن استثنایات به کار گرفته می‌شود، بسته به نوع داشش موجود عمل می‌کند. نقطه شروع مناسب برای توضیح استثنایات در مدل مفهومی انواع مختلف دانشی است که در کاربردهای مختلف استثنایات نقش دارد. دانش ضمنی موجود در سامانه یافتن استثنایات ممکن است بر پایه توضیحی از ساختار نرمال و رفتار وظیفه‌ای سامانه و یا بیان رفتار غیر نرمال سامانه باشد. یافته‌های جمع‌آوری شده‌ای 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 مشاهده شده، 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 نخستین بار توسط ری ریترز (۱۹۸۷) به عنوان چارچوب منطقی برای کشف عارضه‌های سامانه‌های فیزیکی با استفاده از مدل ساختار و رفتار نرمال سامانه مطرح شد. براساس این تئوری، استثنایات براساس مقایسه داده‌های مشاهده شده با ساختار و رفتار نرمال سامانه و مغایرت با آن کشف می‌شوند. کشف استثنایات براساس تطابق با رفتار غیرنرمال با درنظر گرفتن دانش رفتارهای استثنایی سامانه عمل می‌کند، بهنحوی که به شبیه‌سازی رفتار غیر نرمال سامانه می‌پردازد.

به‌منظور ارتقای دقت در کشف استثنایات چارچوب تئوریک تلفیقی جدیدی بر اساس تئوری کشف استثنایات براساس سازگاری و تئوری کشف استثنایات بر اساس تطابق با رفتار غیر نرمال به صورت زیر (شکل ۱) پیشنهاد می‌شود. استثنایات بر مبنای میزان سازگاری داده‌های مشاهده شده با مدل رفتار غیر نرمال و یا مغایرت داده‌های مشاهده شده 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.

طراحی و به کارگیری سامانه خبره بهمنظور استفاده از دانش استخراج شده برای کشف استثنایات جدید.

### ارتقای نظریه استثنایات

چارچوب پژوهش حاضر بدین صورت است: در بخش ۴ در قالب روش‌شناسی پژوهش به بیان نظریه استثنایات پرداخته و سپس مفهوم نمونه‌سازی کاهشی، آنتروپی و چگونگی شناسایی داده‌های استثنایی با به کارگیری آنتروپی رنی عنوان می‌شود. دانش رفتار سهام استثنایی با به کارگیری RISE رویکرد پایین به بالا و براساس روش پیشنهادی 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 به کار گرفته شده و درنهایت نتایج حاصل از پژوهش ارائه می‌شود.

## ۴-روش‌شناسی پژوهش

منطق یادگیری از استثنایات یک مسئله قابل توجه در حوزه یادگیری ماشین است. در پژوهش حاضر مدلی بر اساس رویکرد تلفیقی تئوری استثنایات، تئوری اطلاعات و تئوری دانه‌بندی اطلاعات ارائه شده است تا پدیده‌های استثنایی را کشف کرده و الگوهای رفتاری بنهان آنها را شناسایی کند. ابتدا با هدف کشف استثنایات 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 استثنایات، یادگیری رفتار آنها و به کارگیری قوانین شناسایی شده با به کارگیری تئوری استثنایات و طراحی سامانه خبره تشخیص استثنایات تکمیل می‌شود.

### ۴-نظریه استثنایات

رویکردهای متفاوتی به مسئله استثنایات در حیطه‌های متفاوت علمی و عملی وجود دارد که از آن‌ها می‌توان به استثنایات موضوعی، آماری، ژنتیکی، بیولوژیک و رویکردهای تئوریک اشاره کرد. نخستین بار هافمن (۱۹۷۷) تئوری استثنایات را در علم ژنتیک مطرح کرد. پس از ظهور تئوری استثنایات در علم ژنتیک مک کارتی در سال ۱۹۸۰ مفهوم تئوری استثنایات را در استنتاج مطرح کرد. مک کارتی با به کارگیری قوانین علی و معلولی به شناسایی استثنایات



از داده‌های موجود در پایگاه داده و به کارگیریتابع 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، توسعه استفاده از تئوری اطلاعات به حوزه تحلیل محتوای داده‌ها، برای کشف استثنایات است. نظریه اطلاعات مدلی ریاضی از شرایط و عوامل مؤثر در انتقال و پردازش داده‌ها و اطلاعات با هدف کمی‌سازی و اندازه‌گیری عددی اطلاعات است. نخستین بار مفهوم آنتروپی توسط کلاؤد شانون (۱۹۴۸) در علم فیزیک و برای کمی‌سازی 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 به عنوان ابزاری کارآمد برای سنجش میزان استثنایی بودن داده‌ها استفاده شده است.

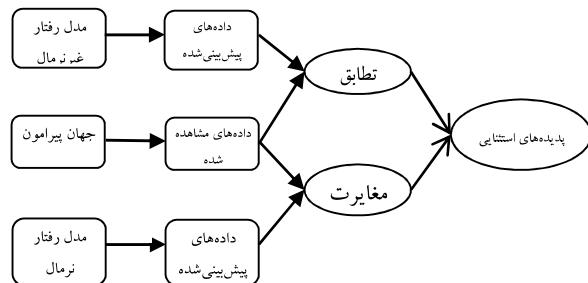
اگر  $X$  یک متغیر تصادفی باشد که یکی از مقادیر  $x_1$  و  $\dots$  و  $x_n$  را با احتمالات  $p_1$  و  $\dots$  و  $p_n$  انتخاب می‌کند، تابع  $H(x)$  نشان‌دهنده میزان آنتروپی حاصل از آن است که میزان ناخالصی برای بروز حالت مشخص  $X$  را نشان می‌دهد. در صورتی که پیشامد  $x_i$  اتفاق بیفتد، با استفاده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

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

از آنجایی که یکی از اهداف این پژوهش، شناسایی داده‌های استثنایی موجود در پایگاه داده است، از تابع آنتروپی رنی به عنوان 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.

$$H(x) = \frac{1}{1-\alpha} \log(\sum p_i^\alpha) \quad (2)$$

<sup>2</sup> Lotfi Zade



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

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

دانه‌بندی اطلاعات به عنوان ابزاری برای محاسبات تخمینی<sup>۱</sup> توسط لطفی‌زاد<sup>۲</sup> پیشنهاد شد. دانه‌بندی اطلاعات یک مجموعه از داده، افزار مجموعه مرجع به  $n$  دانه اطلاعاتی به گونه‌ای است که، داده‌های موجود در یک دانه اطلاعاتی بیشترین شباهت را به یکدیگر داشته باشند. بدین ترتیب یک مسئله دسته‌بندی پیچیده، به چند مسئله ساده‌تر و کوچک‌تر تبدیل شده که منجر به کاهش پیچیدگی فرایند یادگیری می‌شود.

در این رویکرد، فضای مسئله بر اساس میزان شباهت اطلاعات موجود به فضاهای کوچک‌تر افزایش می‌شود. نمونه‌سازی کاهشی با شکستن فضای مسئله، عملکرد مدل طبقه‌بندی کننده را بهبود می‌بخشد. هدف از به کارگیری تئوری دانه‌بندی اطلاعات، افزایش احتمال کشف داده‌های استثنایی و کاهش پیچیدگی مسئله 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 پروسه کشف استثنایات 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 خوشبندی K-means استفاده شده است. این نوع خوشبندی به دلیل استراتژی‌ای که در جداسازی داده‌ها اتخاذ می‌کند، انتخاب شده است و سعی دارد فضای مسئله را به گونه‌ای دسته‌بندی کند که داده‌های هر دسته کمترین فاصله و بیشترین شباهت را به یکدیگر داشته باشند، در حالی که دسته‌های مختلف بیشترین فاصله را از یکدیگر داشته باشند. به دلیل قرارگیری داده‌های مشابه در یک دسته، دقت قوانین استخراج شده از هر دسته افزایش می‌یابد. استثنایات با ساخت دانه‌های اطلاعاتی

<sup>1</sup> Approximate Computing

یادگیری معمول، از رویکرد یادگیری از پایین به بالا در قالب روش پیشنهادی RISE ارتقا یافته، برای کشف دانش رفتار پدیده‌های استثنایی استفاده می‌شود.

رویکرد یادگیری پایین به بالا نخستین بار توسط پدرو دومینگوуз در سال ۱۹۹۸ و در قالب الگوریتم یادگیری RISE ارائه شد. در این روش یادگیری قوانین رفتاری با به کار گیری تمام داده‌های آموزشی صورت می‌پذیرد؛ به طوری که ابتدا مطابق با هر رکورد موجود در پایگاه داده یک ضابطه ساخته شده و سپس با در نظر گیری ضوابط مجاور، تعمیم قوانین صورت می‌پذیرد. این رویکرد تمایل زیادی به کشف قوانین خاص و با دقت بالا دارد؛ اما از آن جایی که هر رکورد موجود در پایگاه داده به عنوان یک ضابطه در نظر گرفته می‌شوند، زمان تعمیم قوانین زیاد است. دومینگوуз روش RISE را به همراه سایر روش‌های موجود یادگیری ماشین مانند الگوریتم‌های درخت تصمیم روی سی پایگاه داده استاندارد به کار 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 به طور مستقیم 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 به خوبی فراگرفته شود.

براساس روش پیشنهادی، پس از شناسایی پدیده‌های استثنایی و تفکیک آن از داده‌های عادی، با درنظر گرفتن چارچوب پیشنهادی تئوری استثنایات، با استخراج قوانین حاکم بر هر دو دسته داده عادی و استثنایی آغاز می‌شود. ابتدا داده‌ها به تفکیک برچسب عادی و استثنایی دسته‌بندی می‌شود. به منظور کاهش هدفمند فضای مسئله و افزایش دقت کشف ضوابط رفتاری، از الگوریتم‌های خوشه‌بندی برای خوشه‌بندی پدیده‌های عادی استفاده می‌شود تا داده‌های نسبتاً مشابه در یک خوشه قرار گیرد. داده‌های استثنایی به دلیل تعداد اندک، نیازی به خوشه‌بندی ندارند. پس از آماده‌سازی خوشه‌های عادی و دسته 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 تشکیل دهنده یک نقطه

از تابع آنتروپی رنسی با  $H(x) = -\sum p_i \log(p_i)$  برای یافتن استثنایات استفاده می‌شود. هر چقدر،  $H(x)$  کوچک‌تر باشد، توانایی تشخیص داده‌های استثنایی بیشتر است. بدین ترتیب که استفاده از توان ۲ برای احتمالات کوچک باعث بزرگ‌تر شدن مقدار تابع  $H(x)$  و وزن دهی بیشتر به آن می‌شود.

$$H(x) = -\sum p_i^2 \quad (3)$$

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

$$H(X, Y) = H(X) + H(Y|X) \quad (4)$$

از آن جایی که  $H(y|X)$  به ازای مقادیر کوچک  $y$  بزرگ‌تر است آنتروپی داده‌های غیرنرمال دارای انحراف زیادی نسبت به متوسط تابع آنتروپی دیگر داده‌ها است.

استثنایات با ساخت دانه‌های اطلاعاتی از داده‌های موجود در پایگاه داده و به کار گیری 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، استثنای خوانده می‌شوند. درنهایت استثنایات موجود در دانه‌های اطلاعاتی تجمعی می‌شوند.

پس از کشف پدیده‌های استثنایی پنهان در پایگاه داده با ایستی 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 استثنایات پرداخته شده است:

#### ۴-۴-استخراج قوانین

اغلب یادگیری رفتار استثنایات فرایند پیچیده‌ای است؛ زیرا رخدادهای استثنایی بسیار کمتر از پدیده‌های معمول اتفاق می‌افتدند. 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 استثنایات است. به منظور غلبه بر کاستی‌های مدل‌های



است که دقت قوانین پیش‌بینی کننده دسته‌های اصلی و فرعی را مورد بررسی قرار می‌دهد.

$$G - \text{means} = \sqrt{\frac{d}{c+d} \cdot \frac{a}{a+b}} \quad (6)$$

با استفاده از معیار ارزیابی G-means و درجه پوشش‌دهی، تغییرات در صورتی 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 خوش‌بندی، تمام متغیرها را در نظر می‌گیرد. به کارگیری استراتژی پوشش متواالی، 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؛ درحالی که 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 استثنایات و یادگیری از آن‌ها مطرح شده است که در ادامه بهتفصیل به گام‌های آن می‌پردازیم:

گام ۱- افزار فضای مسئله بر مبنای میزان شباهت اطلاعاتی داده‌ها با استفاده از روش خوش‌بندی K-means

گام ۲- محاسبه آنتروپی داده‌های موجود در هر خوش

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

گام ۴- تجمعی استثنایات شناخته شده هر خوش

گام ۵- دسته‌بندی داده‌ها- ابتدا داده‌های نرمال و استثنایی شناسایی شده در گام دو را در دو دسته جداگانه قرار داده و سپس داده‌های نرمال دوباره به گونه‌ای

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 نشان‌دهنده ویژگی‌های آن پدیده و متغیر  $y_j = 0, 1$  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:

$$\text{IF } x_1 = a_1, x_2 = a_1, \dots, x_n = a_1 \text{ THEN } y = 1 \quad (5)$$

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

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

هدف از مرحله تعمیم، تعديل ضابطه ایجادشده با حداقل تغییر بهصورتی است که دقت پیش‌بینی در سطح قابل قبولی حفظ شود. در مواجهه با داده جدید، مقدار داده 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 پوشش-دهی، میزان بهبود تعمیم ایجاد شده در ضابطه مورد بررسی ارزیابی می‌شود. در صورتی که 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 نزدیک-ترین 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

تصمیم‌گیری مالی، مبدل کرده است. با این تفاسیر، یافتن مجموعه 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  
(Table-1): Rule extraction based on E-Rise algorithm

مرحله ایجاد قوانین	مرحله عمومی‌سازی
ES is the training set SS= select $\alpha$ % of ES randomly Let RS be SS For each rule R in RS N= the nearest neighborhoods E to R by $d < d^*$ let $\tilde{R}$ = Generalization (R,N) Let RS = RS with R replaced R with $\tilde{R}$ if Acc( $\tilde{R}$ ) $\geq \beta\% Acc(RS)$ Then replace RS by $\tilde{R}$ if $\tilde{R}$ is identical to another rule in RS then delete $\tilde{R}$ from RS Until $Acc(RS) \geq \gamma$ Return RS	$R =$ $a_1, a_2, \dots, a_m, C_R$ $N =$ $e_1, e_2, \dots, e_m, C_n$ نمونه نزدیک باشد : function generalization (R,N) $a_i$ is either true , $x_i = r_i$ or $r_{i,lower} \leq x_i \leq r_{i,upper}$ 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 $e_i = r_{i,lower}$

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 به صورت ماهیانه، از نیمه دوم سال ۱۳۹۰ تا پایان سال ۱۳۹۲ در ۶۰۲۲ رکورد، گردآوری شده است. بهمنظور تشکیل پروفایل از سهام مورد بررسی و کشف استثنایات، لازم است مجموعه قابل نمایشی از سهام ایجاد شود.

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

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

گام ۷- قوانین ایجاد شده با در نظر گیری 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 درصورتی لحاظ می‌شوند که شاخص G-means بدتر نشده و قانون ایجادشده درصد مشخصی از داده‌های موجود در آن دسته را پوشاند.

گام ۸- طراحی سیستم خبره برای شناسایی استثنایات جدید با استفاده از چارچوب پیشنهادی تئوری استثنایات. استثنایات 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، به عنوان پدیده‌های استثنایی تلقی می‌شوند. شکل (۲) فرایند کشف و یادگیری از رفتار استثنایات را نشان می‌دهد.

## ۵- یافته‌های پژوهش

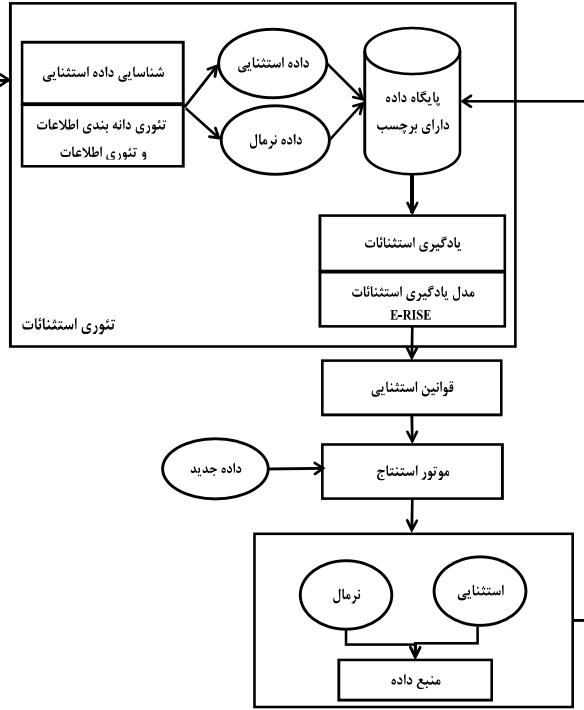
در پژوهش حاضر، کشف پدیده‌های استثنایی و یادگیری رفتار آن‌ها در قالب مدل پیشنهادی و به‌منظور افزایش اطمینان به پدیده‌های استثنایی، انجام می‌پذیرد. کارایی الگوی پیشنهادی با مطالعه رفتار سهام موجود در بازار اوراق بهادار ایران با هدف شناسایی سهام استثنایی و یادگیری رفتار آن‌ها، بررسی شده است. بازار سهام 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 جدی مواجه ساخته است؛ زیرا به کارگیری عوامل مختلف باعث ایجاد نتایج متفاوت می‌شود. از سوی دیگر، مسئله عدم شفافیت و قطعیت در بازار سهام، انتخاب سبد سهام استثنایی را به یک معضل اساسی در حوزه



مرحله پیش‌پردازش داده‌ها انجام شده است. به این منظور، ابتدا مغایرت‌های موجود در پایگاه داده، شناسایی و بطرف شد. اغلب، منشأ این مغایرت‌ها منابع مختلف گردآوری اطلاعات است. به منظور پاکسازی داده‌ها از عادی‌سازی، شناخت نقاط دور افتاده و پرکردن داده‌های مفقود استفاده شده است. به دلیل تعدد منابع داده، درجه ناهمنگ بودن و آلودگی داده‌ها، اجرای چندین گام عمل تبدیل و پاکسازی داده‌ها ضروری بوده است.

برای کشف سهام استثنایی، ابتدا با استفاده از "نمونه‌سازی کاهشی"، دانه‌های اطلاعاتی ساخته می‌شود. روش خوشبندی K-means برای تشکیل دانه‌های اطلاعاتی به کار گرفته شده و درنتیجه، سه خوشه ایجاد شده است. بدین ترتیب، فضای مسئله بر مبنای میزان شباهت اطلاعاتی داده‌ها، به فضاهای کوچک‌تر، افزایش می‌شود که باعث بهبود عملکرد مدل طبقه‌بندی خواهد شد؛ سپس، ازتابع آنتروپی رنی برای تشخیص بینظمی‌های هر دانه اطلاعاتی استفاده می‌شود. براساس تابع آنتروپی، فضای هر دانه اطلاعاتی به دو دسته عادی و استثنایی تقسیم شده است.

در پژوهش حاضر، سهام استثنایی سهامی هستند که میزان تابع آنتروپی آنها به فاصله  $1/55 + 1$  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.



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

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

فرض کنید  $x_i$  نشان‌دهنده سهم  $i$  در مشخصه  $j$  باشد، در حالی که  $\{x_{1,1}, x_{1,2}, \dots, x_{1,7}\} = i$  نشان‌دهنده ۶۰۲۲ سهم مورد بررسی و  $\{x_{6022,1}, x_{6022,2}, \dots, x_{6022,7}\} = j$  متغیرهای درنظر گرفته شده در شناسایی و یادگیری رفتار استثنایی سهام است. در این صورت  $M$  نشان‌گر پروفایل سهام مورد بررسی خواهد بود.

$$M = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,7} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,7} \\ \vdots & \vdots & \cdots & \vdots \\ x_{6022,1} & x_{6022,2} & \cdots & x_{6022,7} \end{bmatrix} \quad (1-4)$$

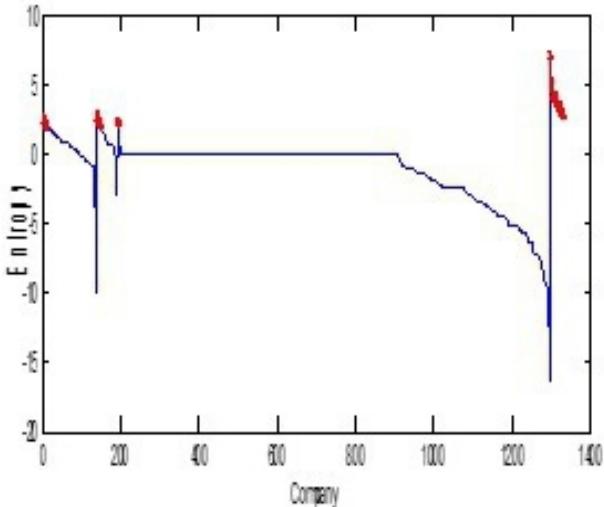
به دلیل اهمیت دقیق بودن ورودی‌های مدل و تأثیر آلودگی داده‌ها بر تنزل عملکرد فرایند داده کاوی،

(جدول-۲): مشخصه سهام مورد بررسی

(Table-2): stocks characteristics

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

درصدی ( $\alpha\%$ ) 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  $\alpha$  در خوشه‌های مختلف، لحاظ شده است. جدول (۴) تأثیر تغییر در مقدار  $\alpha$  را بر دقّت پیش‌بینی و تعداد قانون استخراج شده نشان می‌دهد. با توجه به اطلاعات جدول (۳) بهترین مقدار برای  $\alpha$  در خوشه‌های نخست و دوم و سوم و چهارم به ترتیب برابر با  $1/5$ ،  $1/4$ ،  $1/3$  و  $1/5$ % 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 به کارگیری داده‌های مجاور ضابطه اولیه، که تا به حال 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. درصورتی که اندازه فاصله کم باشد، تعداد تکرارها و زمان تعمیم افزایش می‌یابد و درصورتی که فاصله زیاد باشد، دقّت قوانین کاهش خواهد یافت. با درنظر گرفتن فاصله همسایگی مناسب، الگوی رفتاری استثنائات در قالب سه ضابطه موجود در جدول (۵) شناسایی شده است.



(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 خوشبندی شده است که شامل یک خوشه از داده‌های استثنایی و سه خوشه از داده‌های عادی می‌شود. استفاده از خوشبندی 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.

(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 ۲- 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 مستخراج  
(Table-4): optimal number of extracted rules

$\alpha_8$	$\alpha_7$	$\alpha_6$	$\alpha_5$	$\alpha_4$	$\alpha_3$	$\alpha_2$	$\alpha_1$	تعداد داده	خوش
0.4	0.3	0.3	0.3	0.4	0.4	0.4	0.5	۹	۱
0.0015	0.002	0.003	0.004	0.005	0.006	0.007	0.008	۴۶۲۸	۲
0.004	0.005	0.005	0.006	0.006	0.008	0.01	0.01	۸۲۷	۳
0.005	0.005	0.01	0.01	0.02	0.02	0.03	0.03	۵۲۳	۴
16	18	25	31	42	48	59	65	تعداد قانون استخراجی	
86.53	83.74	70.09	64.87	59.87	60.37	53.46	52.35	G-means (%)	

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

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

آنگاه سهام	آگر
استثنایی	قیمت به سود < ۶/۴۶ و بازده < ۴۴/۵۹ و حجم معاملات < ۹۴۶۵۰۰ و ارزش معاملات < ۵۵۶ و قیمت < ۱۷۰۲۵
استثنایی	بازده < ۰/۶۴ و ارزش معاملات < ۹۹۱۵۴۴۵۵۷۶ و حجم معاملات < ۹۸۰۲۴۰۹۴۹۶۴۳۳۶
استثنایی	بازده < ۰/۲۶ و ارزش معاملات < ۴۹۲۰۸ و قیمت < ۲۷۴۸۷۹۵۶۲

الگوریتم دسته‌بندی CHAID تمامی داده‌ها را به عنوان داده عادی شناسایی و قوانین رفتاری آن‌ها را کشف کرده است. الگوریتم CART تنها ۳۵ پدیده استثنایی را به درستی پیش‌بینی کرد در حالی که از لحاظ دقت پیش‌بینی ۹۷/۵۶ درصد از پدیده‌ها را به درستی 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 به درستی شناسایی کرده است؛ اما برای ایجاد قوانین فقط از سه متغیر نسبت قیمت به سود، دفعات معاملات و قیمت استفاده کرده است.

خروچی‌های حاصل از مدل‌های به کار گرفته شده، نشان می‌دهد این الگوریتم‌ها برخلاف دقت بالا در پیش‌بینی، به دلیل 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 این‌وه داده موجود در پایگاه داده و یادگیری رفتار آن‌ها است. در پژوهش حاضر، رویکرد تدوین مجدد مسئله اصلی برگزیده شده است؛ به این

با استخراج قوانین حاکم بر رفتار عادی و استثنایی سهام، به منظور پیش‌بینی وضعیت سهام جدید، سامانه خبره شناسایی سهام استثنایی در محیط نرم‌افزار GURU طراحی شده است. این سامانه با استفاده از قوانین شناسایی شده 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 خبره طراحی شده به منظور کشف سهام استثنایی موجود در پایگاه داده سهام، به کار گرفته شده است تا قابلیت قواعد استخراج شده توسط الگوریتم RISE ارتقایافته، سنجیده شود. نتایج حاصل از به کار گیری سامانه خبره طراحی شده روی پایگاه داده سهام، نشان دهنده قابلیت روش پیشنهادی در کشف سهام استثنایی و یادگیری رفتار آن‌هاست.

به منظور ارزیابی عملکرد مدل پیشنهادی کشف و یادگیری از استثنایات، عملکرد الگوریتم‌های یادگیری C4.5، CART و CHAID روى پایگاه داده پاکسازی شده با استفاده از نرم‌افزار کلمانتاین 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.

## 7-References

## ۷-مراجع

- [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. Icz, "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.

ترتیب که استثنایات شناسایی می‌شود و در یک دسته جدآگانه قرار می‌گیرد؛ سپس، با به کارگیری الگوریتم E-RISE، قوانین رفتار عادی و استثنایی استخراج می‌شود. این روش یادگیری، به دلیل عدم شکست فضای مسأله و به کارگیری تمام متغیرها در فرایند یادگیری، مانع از دست رفتن اطلاعات می‌شود. همچنین فرایند عمومی‌سازی در هر بار برای یک متغیر خاص، انجام نمی‌شود؛ بلکه این روش، تمام متغیرها را در نظر می‌گیرد؛ بدین صورت که انتخاب داده‌نژدیک از بین تمام نقاط موجود در یک دسته، انجام می‌پذیرد.

به منظور 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 بازه زمانی ۱۳۹۰-۱۳۹۲ در هفت متغیر بازده، نسبت قیمت به سود، حجم معاملات، دفعات معاملات، ارزش معاملات، کمترین قیمت و بیشترین قیمت مورد بررسی قرار گرفت. از قوانین استخراجی برای پیش‌بینی رفتار سهام جدید و تشکیل پرتفوی بهینه استفاده می‌شود. پیشنهاد می‌شود به منظور افزایش دقت تشخیص آستانه رفتار استثنایی سهام از منطق فازی در تعیین آستانه‌های قابل قبول بروز رفتار استثنایی به کار گرفته شود که می‌تواند به افزایش دقت فرایند شناسایی استثنایات منجر شود.

(جدول-۶): ماتریس درهم‌بختگی

(Table-6): confusion matrix

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



- 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 حاجی گل یزدی** دانش آموخته  
قطع 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 رایانمه ایشان عبارت است از:  
**elahehajigol@gmail.com**
- مسعود عابسی** دانش آموخته  
دانشگاه ایالتی کلمنسن 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  
تریبیت مدرس، و مشاور داده کاوی مدیر عامل سازمان بیمه  
سلامت ایران است. وی قبلًا مشاور داده کاوی مدیر عامل  
بانک اقتصاد نوین، سازمان پشتیبان تأمین اجتماعی،  
شرکت سوخت رسانی هواپیمایی اوج و مشاور مدیر عامل  
شرکت فولاد آلیاژی ایران بوده است. نامبرده استاد مدعو  
دانشگاه دولتی مالزی، دانشگاه تهران و دانشگاه اسکرانتون  
ایالت پنسیلوانیا آمریکا بوده است.  
نشانی رایانمه ایشان عبارت است از:  
**Mabessi@yazd.ac.ir**
- محمد باقر فخرزاد** دانش آموخته  
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 Cities, 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



**elahehajigol@gmail.com**



**Mabessi@yazd.ac.ir**



سال ۱۳۹۶ شماره ۱ پیاپی ۲۱

و داده کاوی فعالیت می کند.

نشانی رایانمہ ایشان عبارت است از:

[mfakhrzad@yazd.ac.ir](mailto:mfakhrzad@yazd.ac.ir)

### حسن حسینی نسب داش آموخته



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 برنامه ریزی تولید، طراحی سامانه های تولیدی، سامانه های ساخت و تولید پیشرفته و برنامه ریزی استراتژیک فعالیت می کند.

نشانی رایانمہ ایشان عبارت است از:

[hhn@yazduni.ac.ir](mailto:hhn@yazduni.ac.ir)