



تدوین استراتژی معاملاتی بر اساس خوشه بندی داده های قیمت با الگوریتم ماشین لرنینگ DBSCAN و یافتن داده های نویز

تاریخ دریافت مقاله: ۱۴۰۳/۰۹/۲۱ تاریخ پذیرش مقاله:

نسترن شریفی راد^۱

محمد رضا رستمی^۲

چکیده

در این پژوهش، یک استراتژی معاملاتی مبتنی بر خوشه بندی داده های قیمتی با استفاده از الگوریتم ماشین لرنینگ DBSCAN ارائه و عملکرد آن در بازار بورس تهران ارزیابی شده است. برای این منظور، داده های تاریخی نمادهای معاملاتی در بازار بورس مورد بررسی قرار گرفت و ویژگی های کلیدی از داده های قیمت و حجم استخراج شده است. عملکرد این استراتژی با استفاده از شاخص بازدهی استراتژی ها مقایسه شد و نتایج آن با یک روش آماری کلاسیک (بر اساس توزیع نرمال و تحلیل انحراف معیار) و استراتژی سنتی خرید و نگهداری مقایسه گردید.

نتایج نشان داد که استراتژی مبتنی بر DBSCAN، با توانایی شناسایی خوشه های پیچیده و داده های پرت، عملکرد بهتری نسبت به روش آماری داشته و بازدهی بیشتری ارائه کرده است. همچنین، این استراتژی در مقایسه با استراتژی خرید و نگهداری نه تنها بازدهی بالاتری داشت، بلکه توانست ریسک معاملاتی را در بازارهای رنج و نزولی کاهش دهد. یافته های این تحقیق نشان می دهد که استفاده از الگوریتم های پیشرفته خوشه بندی می تواند ابزار موثری برای تحلیل بازارهای مالی و تصمیم گیری معاملاتی در بازار بورس تهران فراهم کند.

کلمات کلیدی

استراتژی معاملاتی، ماشین لرنینگ، خوشه بندی، داده های پرت

۱- دانشجوی کارشناسی ارشد، گروه مدیریت مالی، واحد تهران، دانشگاه الزهراء، ایران. N.sharifirad@student.alzahra.ac.ir

۲- دانشیار، گروه مدیریت مالی، واحد تهران، دانشگاه الزهراء، ایران. M.rostami@alzahra.ac.ir

مقدمه

بازار سهام، میدانی پویا و پر از نوسانات است که پیش‌بینی روند آن همواره برای سرمایه‌گذاران جذاب بوده است. با این حال، طبق نظریه بازار کارآمد، پیش‌بینی دقیق و مستمر قیمت سهام تقریباً غیرممکن است. این نظریه بیان می‌کند که قیمت سهام بازتاب‌دهنده تمام اطلاعات موجود در بازار است و هرگونه تغییر قیمتی به صورت تصادفی رخ می‌دهد. اما واقعیت بازارهای مالی نشان می‌دهد که همیشه استثناهایی وجود دارد. برخی از بازارها، به ویژه در شرایط خاص، از کارآمدی کامل فاصله می‌گیرند. در چنین شرایطی، داده‌های قیمت سهام دیگر کاملاً تصادفی نبوده و الگوهایی در آن‌ها قابل شناسایی است. یکی از چالش‌های اصلی در پیش‌بینی روند سهام، وجود نویز^۱های فراوان در داده‌های قیمت است. این نویزها ناشی از عوامل مختلفی مانند شایعات، اخبار کوتاه‌مدت و تصمیمات لحظه‌ای سرمایه‌گذاران است. روش‌های سنتی پیش‌بینی که بر اساس داده‌های قیمت استوارند، اغلب تحت تأثیر این نویزها قرار گرفته و دقت پیش‌بینی کمی دارند. علاوه بر نویز، پدیده‌هایی مانند معاملات داخلی و دستکاری بازار نیز بر روند قیمت سهام تأثیر می‌گذارند. این فعالیت‌ها که اغلب بر اساس اطلاعات غیر عمومی یا با هدف گمراه کردن سرمایه‌گذاران انجام می‌شوند، باعث ایجاد ناهنجاری‌هایی در حجم معاملات می‌شوند (نیپان^۲ و همکاران، ۲۰۲۱).

حجم معاملات نشان‌دهنده میزان فعالیت معاملاتی در یک بازار است و می‌تواند اطلاعات ارزشمندی در مورد انتظارات بازار و تغییرات قیمت در آینده ارائه دهد. مطالعات بسیاری به بررسی رابطه بین حجم معاملات و بازده سهام پرداخته‌اند و نتایج متفاوتی را گزارش کرده‌اند. برخی مطالعات نشان داده‌اند که رابطه‌ی مثبت و معناداری بین حجم معاملات و بازده سهام وجود دارد، در حالی که برخی دیگر رابطه‌ی منفی یا عدم رابطه‌ی معنی‌داری را گزارش کرده‌اند. این تفاوت در نتایج می‌تواند به عوامل مختلفی مانند نوع بازار، دوره زمانی، فرکانس داده‌ها و روش‌های اندازه‌گیری بستگی داشته باشد (نیپان و همکاران، ۲۰۲۱).

علاوه بر این، مطالعات اخیر نشان داده‌اند که حجم معاملات می‌تواند بر گشتاورهای مرتبه بالاتر بازده سهام، مانند چولگی^۳ و کشیدگی^۴، نیز تأثیر بگذارد. چولگی نشان‌دهنده عدم تقارن توزیع بازده‌ها و کشیدگی نشان‌دهنده ضخامت دم‌های توزیع است. این گشتاورها می‌توانند اطلاعات مهمی در مورد ریسک و بازده بالقوه سرمایه‌گذاری‌ها ارائه دهند. با این حال، در بسیاری از مطالعات، به تأثیر داده‌های پرت^۵، بر رابطه‌ی بین حجم معاملات و نوسانات بازده توجه کافی نشده است. داده‌های پرت مشاهداتی هستند که با الگوی کلی داده‌ها سازگار نیستند و می‌توانند ناشی از خطاهای اندازه‌گیری یا رویدادهای

استراتژی معاملاتی بر اساس خوشه بندی داده های قیمت.../ شریفی، رستمی

غیرمعمول باشند. وجود داده های پرت می تواند منجر به تخمین های نادرست و نتایج گمراه کننده شود (احدزی و جیاسریدهران^۶، ۲۰۲۰).

داده های پرت یا غیرعادی می توانند نتایج تحلیل های آماری را مخدوش کنند، اما در عین حال حاوی اطلاعات ارزشمندی نیز هستند. این موضوع در حوزه مالی و اقتصادی و به ویژه در تحلیل بازده سهام اهمیت بسزایی دارد. مثلاً وجود داده های پرت در بازده، که ناشی از تغییرات شدید قیمت هستند، می توانند بر تصمیمات سرمایه گذاری و عملکرد سبد سهام تأثیر بگذارند (شهاده^۷ و همکاران، ۲۰۲۲).

بررسی داده های پرت اهمیت شناسایی و تحلیل داده های پرت در بازارهای سهام را برجسته می کند و تأثیر آن ها را بر مدل سازی مالی و اقتصادی، مدیریت ریسک، تصمیم گیری سرمایه گذاری و سیاست گذاری اقتصادی مورد بحث قرار می دهد. همچنین داده های پرت تأثیر قابل توجهی بر عملکرد و تحقق بازده سهام دارند و می توانند بر تصمیمات تخصیص دارایی ها و مدیریت ریسک تأثیر بگذارند. این نوع داده ها اطلاعات مهمی در مورد پویایی و رفتار بازار ارائه دهند که به نوبه خود می تواند منجر به تصمیم گیری های قابل اعتمادتری در مورد مدیریت ریسک و سبد سهام شود (شهاده و همکاران، ۲۰۲۲).

در بازارهای مالی، وجود داده های پرت می تواند تأثیرات عمیقی بر تحلیل ها و مدل سازی ها داشته باشد. این داده های غیرعادی، که ممکن است ناشی از خطاهای اندازه گیری یا رویدادهای غیرمنتظره باشند، می توانند باعث انحراف در نتایج و استنتاج های آماری شوند. یکی از حوزه هایی که داده های پرت در آن اهمیت ویژه ای دارند، تحلیل سری های زمانی مالی و اقتصادی است. این سری های زمانی اغلب دارای توزیع هایی با دم های چاق هستند که احتمال وقوع مقادیر پرت در آن ها بالاتر است (گونزالس-سانچز^۸، ۲۰۲۱).

یکی از نمونه های بارز داده های پرت در بازارهای مالی، سقوط های ناگهانی بازار است که می توانند تأثیرات شدیدی بر سرمایه گذاران و کل اقتصاد داشته باشند (شلتز و ترد^۹، ۲۰۰۸). شناسایی و درک این داده های پرت برای مدیریت ریسک، پیش بینی نوسانات بازار و تصمیم گیری های سرمایه گذاری حیاتی است. در این مقاله، ما به بررسی داده های پرت در حجم معاملات و قیمت دارایی ها در بازار سهام می پردازیم و تأثیر آن ها را بر پیش بینی روند سهام بررسی می کنیم.

در این راستا، ما از داده های بازار سهام استفاده خواهیم کرد و با بهره گیری از یک الگوریتم خوشه بندی یادگیری ماشین^{۱۰} روش های آماری و تکنیک های داده کاوی، به شناسایی و تحلیل داده های پرت در حجم

معاملات خواهیم پرداخت. سپس، با بررسی ارتباط بین این داده‌های پرت و تغییرات قیمت سهام، سعی خواهیم کرد تا میزان قابلیت پیش‌بینی روند سهام را با استفاده از این اطلاعات ارزیابی کنیم. هدف ما از این مطالعه، ارائه یک تحلیل جامع از رابطه‌ی بین حجم معاملات و نوسانات بازده در بازار سهام است. همچنین به جای تمرکز صرف بر داده‌های قیمت، ما به دنبال شناسایی و بهره‌برداری از ناهنجاری‌هایی هستیم که در حجم معاملات رخ می‌دهند. این ناهنجاری‌ها که ناشی از معاملات داخلی و دستکاری بازار هستند، می‌توانند نشانه‌های مهمی از تغییرات آتی قیمت باشند. با شناسایی و تحلیل این ناهنجاری‌ها، می‌توانیم تأثیر نویزهای قیمت را بررسی کرد و پیش‌بینی دقیق‌تری از روند قیمتی سهام ارائه دهیم.

یافته‌های این پژوهش می‌تواند به سرمایه‌گذاران، تحلیلگران و سایر فعالان بازار کمک کند تا درک بهتری از نقش داده‌های پرت در حجم معاملات و تأثیر آن‌ها بر روند سهام داشته باشند و در نتیجه، تصمیمات سرمایه‌گذاری آگاهانه‌تری اتخاذ کنند. همچنین، این مطالعه می‌تواند به توسعه مدل‌های پیش‌بینی روند سهام و بهبود مدیریت ریسک در بازارهای مالی کمک کند. ما امیدواریم که یافته‌های این مطالعه بتواند به درک بهتر رفتار بازار و توسعه استراتژی‌های معاملاتی و مدیریت ریسک کمک کند. همچنین، این مطالعه می‌تواند برای محققانی که در زمینه‌های مرتبط با بازارهای مالی فعالیت می‌کنند، مفید باشد.

برای درک بهتر اهمیت شناسایی داده‌های پرت و کاربرد آن در بازارهای مالی، مروری بر مطالعات پیشین انجام شده در این زمینه ضروری است. همچنین در ادامه، به بررسی چند نمونه از پژوهش‌های اخیر که به طور خاص به استفاده از الگوریتم خوشه‌بندی مبتنی بر چگالی با وجود نویز^{۱۱} برای شناسایی داده‌های پرت پرداخته‌اند، اشاره خواهیم کرد. این مطالعات، زمینه‌های مختلف کاربرد این الگوریتم، از جمله خوشه بندی داده‌ها با الگوهای پیچیده و توانایی تشخیص و مدیریت داده‌های پرت توسط این الگوریتم را پوشش می‌دهند.

در پژوهش نخست به بررسی تأثیر شکست‌ها و داده‌های پرت بر مدل‌سازی نوسانات و پایداری نوسانات در بازار سهام ایالات متحده می‌پردازد. برای این منظور، از یک الگوریتم مبتنی بر موجک برای شناسایی داده‌های پرت و یک روش آماری جمع انباشته^{۱۲} برای شناسایی شکست‌ها در سری‌های زمانی بازده سهام استفاده می‌شود. نتایج نشان می‌دهد که وجود داده‌های پرت و شکست‌ها می‌تواند به طور قابل توجهی بر تخمین پایداری نوسانات تأثیر بگذارد. به طور خاص، نادیده گرفتن شکست‌ها می‌تواند منجر به برآورد بیش از حد پایداری نوسانات شود، در حالی که نادیده گرفتن داده‌های پرت می‌تواند منجر به سوگیری در تخمین پارامترهای مدل شود. این یافته‌ها پیامدهای مهمی برای مدل‌سازی و پیش‌بینی نوسانات در

استراتژی معاملاتی بر اساس خوشه بندی داده های قیمت.../ شریفی، رستمی

بازارهای مالی دارند. در نهایت، این مطالعه نشان می‌دهد که در نظر گرفتن شکست‌ها و داده‌های پرت در مدل‌سازی نوسانات می‌تواند منجر به بهبود قابل‌توجهی در دقت مدل‌ها شود. این امر به‌ویژه در بازارهای مالی که در معرض شوک‌ها و رویدادهای غیرمنتظره هستند، اهمیت دارد. (چاتزیکنستانتی^{۱۳}، ۲۰۱۷)

این پژوهش با استفاده از روش نمودار جعبه‌ای به شناسایی و تحلیل داده‌های پرت در بازه روزانه ۱۴ شاخص بازار سهام بین‌المللی پرداخته است. نتایج نشان می‌دهد که حدود ۶ درصد از مشاهدات داده‌های پرت هستند، با ۱.۴ درصد داده‌های پرت شدید. داده‌های پرت منفی تأثیرگذارتر و شدیدتر بوده و بیشتر در حدود سه انحراف استاندارد مشاهده می‌شوند. همچنین، این داده‌ها تمایل به خوشه‌بندی دارند، چه در سری‌های زمانی فردی و چه بین بازارها. این یافته‌ها ملاحظات مهمی برای سرمایه‌گذاری، مدیریت پرتفوی و ریسک ارائه می‌دهند، زیرا داده‌های پرت، با وجود سهم کم، تأثیر قابل‌توجهی بر عملکرد بازار سهام دارند. (شهاده و همکاران، ۲۰۲۲).

در این پژوهش، یک روش جدید داده‌کاوی برای پیش‌بینی رفتار بلندمدت روند سهام ارائه شده است. در این روش، با تمرکز بر روی توزیع حجم در داده‌های معاملاتی با فرکانس بالا^{۱۴}، ناهنجاری‌ها در توزیع حجم معاملات را شناسایی می‌کند و به‌عنوان نشانه‌ای از ناکارآمدی بازار در نظر گرفته می‌شوند. الگوریتم پیشنهادی با تحلیل داده‌های تیک-به-تیک و شناسایی ناهنجاری‌ها، روند سهام را با استفاده از خوشه‌بندی پیش‌بینی می‌کند. نتایج در بازار سهام چین نشان می‌دهد که این روش در پیش‌بینی بلندمدت سودآور بوده است (ژائو و وانگ، ۲۰۱۵).

این پژوهش یک روش آماری جدید برای شناسایی داده‌های پرت در توزیع‌های دم‌کلفت^{۱۵} ارائه می‌دهد و آن را در بررسی سقوط بازار سهام سال ۱۹۸۷ به کار می‌گیرد. روش پیشنهادی مبتنی بر آزمون نسبت آماره‌های ترتیبی متوالی است و برای شناسایی چندین داده پرت طراحی شده است. با اعمال این روش بر روی داده‌های بازده شاخص داو جونز، پس از حذف وابستگی زمانی، نتیجه گرفته می‌شود که سقوط بازار در سال ۱۹۸۷ یک داده پرت نبوده و با توزیع معمول بازده‌ها سازگار است. این یافته‌ها اهمیت در نظر گرفتن داده‌های پرت و شکست‌های ساختاری در مدل‌سازی و پیش‌بینی نوسانات بازارهای مالی را برجسته می‌سازد. (شلتر و ترد، ۲۰۰۸).

این پژوهش رابطه بین حجم معاملات و بازده سهام در بازار بورس مصر را بررسی می‌کند. با استفاده از لگاریتم نسبت گردش مالی و تعداد معاملات برای اندازه‌گیری حجم، نتایج نشان‌دهنده رابطه مثبت و معنادار بین حجم معاملات و بازده سهام است. همچنین، رابطه همزمانی مثبت اما ضعیف و رابطه منفی

با وقفه مشاهده شد. آزمون علیت گرنجر نشان‌دهنده رابطه علی دوطرفه بین حجم معاملات و بازده، به‌ویژه با داده‌های پنج روز گذشته است. این مطالعه نشان می‌دهد که تعداد معاملات معیار دقیق‌تری برای حجم معاملات نسبت به لگاریتم نسبت گردش مالی است و به درک ساختار و ناکارآمدی اطلاعاتی بازار بورس مصر کمک می‌کند. (الانصاری و آتوئه^{۱۶}، ۲۰۱۲).

این پژوهش یک روش بدون نظارت برای تشخیص داده‌های پرت با استفاده از الگوریتم اصلاح‌شده خوشه بندی میانگین^{۱۷} ارائه می‌دهد. داده‌های پرت شناسایی شده از مجموعه داده حذف می‌شوند تا دقت خوشه‌بندی بهبود یابد. رویکرد پیشنهادی با مقایسه با تکنیک‌های موجود و عملکرد معیار اعتبارسنجی می‌شود. نتایج تجربی روی مجموعه داده‌های معیار نشان می‌دهد که تکنیک پیشنهادی از روش‌های موجود در چندین معیار بهتر عمل می‌کند. (احمد و محمود^{۱۸}، ۲۰۱۳)

این پژوهش رابطه بین بازده سهام و حجم معاملات در بازار بورس کراچی پاکستان را با استفاده از داده‌های هفتگی شاخص اصلی بورس از ژانویه ۲۰۰۰ تا آوریل ۲۰۱۲ بررسی می‌کند. نتایج نشان می‌دهد که بین بازده و حجم معاملات، همچنین بین بازده و تغییرات حجم معاملات، رابطه مثبت و معناداری وجود دارد. آزمون علیت گرنجر نیز رابطه دوطرفه بین این دو متغیر را تأیید می‌کند. این یافته‌ها نشان می‌دهد که حجم معاملات اطلاعاتی درباره انتظارات بازار و تغییرات قیمت ارائه می‌دهد و می‌تواند برای سرمایه‌گذاران و سیاست‌گذاران مفید باشد (عطاری^{۱۹} و همکاران، ۲۰۱۲).

این پژوهش به بررسی روش‌های تحلیل حجم معاملات و کاربرد آنها در پیش‌بینی حرکات قیمت در بازارهای مالی می‌پردازد و با تمرکز بر اهمیت تحلیل حجم معاملات، روش‌ها و رویکردهای مختلفی مانند پروفایل بازار و پروفایل حجم را برای پیش‌بینی قیمت‌ها و به حداکثر رساندن سود معاملاتی ارائه می‌دهد. در نهایت، این مقاله نشان می‌دهد که تحلیل حجم معاملات می‌تواند به عنوان یک ابزار مفید برای پیش‌بینی حرکات قیمت در بازارهای مالی مورد استفاده قرار گیرد، اما استفاده موثر از آن نیازمند دانش و تجربه کافی در تحلیل بازار است و معامله‌گران باید به محدودیت‌ها و خطرات مرتبط با این روش‌ها توجه کنند. (چوتکا و ربتاک^{۲۰}، ۲۰۲۱).

این مطالعه رابطه بین حجم معاملات و شاخص‌های سهام را بررسی کرده و نشان می‌دهد حجم معاملات می‌تواند برای پیش‌بینی حرکات شاخص‌ها مفید باشد. با تحلیل داده‌های روزانه (۲۰۰۰ تا ۲۰۱۰)، نتایج حاکی از آن است که آگاهی از حجم معاملات هفته گذشته به پیش‌بینی سطح شاخص‌ها کمک می‌کند، به‌ویژه در بازه‌های طولانی‌تر مانند ۱۶ ماهه (۲۰۰۹-۲۰۱۰). یافته‌ها نشان می‌دهد تحلیل تکنیکال

استراتژی معاملاتی بر اساس خوشه بندی داده های قیمت.../ شریفی، رستمی

می تواند برای سرمایه گذاری بلندمدت نیز مؤثر باشد. این پژوهش پیشنهاد می کند تحقیقات آینده به پیش بینی خارج از نمونه و تعیین تأخیر بهینه برای هر شاخص بپردازد (آبوندانته^{۲۱}، ۲۰۱۰).

مقاله آخر نیز یک مدل بازار سهام را بررسی می کند که در آن در دسترس بودن اطلاعات خصوصی به صورت تصادفی در طول زمان تغییر می کند. در این مدل، معامله گران از حجم معاملات گذشته برای تخمین در دسترس بودن اطلاعات خصوصی استفاده می کنند. با افزایش احتمال وجود اطلاعات خصوصی، معامله گران محتاط تر می شوند و سفارش های محدود محافظه کارانه تری ثبت می کنند. این مدل با الگوهای تجربی مشاهده شده سازگار است و همبستگی مثبت بین تغییرپذیری قیمت و حجم معاملات را نشان می دهد. همچنین، واریانس شرطی خودهمبسته و میانگین برگشتی است و ممکن است با حجم معاملات مورد انتظار همبستگی مثبت یا منفی داشته باشد. در نهایت، تغییرات قیمت برای مشخص کردن تکامل واریانس شرطی کافی نیست و اطلاعات مربوط به حجم معاملات نیز مورد نیاز است. (سومینن^{۲۲}، ۲۰۰۱)

همچنین در ادامه مقالاتی در جهت استفاده از روش های ماشین لرنینگ مثل نزدیکترین همسایه و خوشه بندی مبتنی بر چگالی پیدا کردن خوشه ها و داده های پرت معرفی می گردد.

این پژوهش رویکرد جدیدی برای شناسایی داده های پرت با استفاده از الگوریتم خوشه بندی میانگین اصلاح شده ارائه می دهد. داده های پرت شناسایی شده حذف می شوند تا دقت خوشه بندی بهبود یابد. نتایج تجربی نشان می دهد این روش از تکنیک های موجود بهتر عمل می کند و کاربردهای مهمی در شناسایی بیماری، تشخیص نفوذ شبکه، کشف تقلب مالی و تحلیل رفتار انسانی دارد. پژوهش به محدودیت روش های موجود، مانند پیچیدگی محاسباتی و دقت پایین، اشاره می کند و الگوریتم های خوشه بندی مقاوم مانند خوشه بندی قوی با استفاده از پیوندها^{۲۳} و روش های مبتنی بر چگالی را بررسی می کند که توانایی مدیریت داده های پرت را نیز دارند. (احمد و محمود، ۲۰۱۳)

این پژوهش یک الگوریتم خوشه بندی مبتنی بر چگالی اصلاح شده را برای خوشه بندی داده ها با خوشه های مجاور متراکم ارائه می دهد. الگوریتم خوشه بندی مبتنی بر چگالی اصلی در تشخیص اشیاء مرزی خوشه های مجاور است و نتیجه خوشه بندی به ترتیب پردازش اشیاء بستگی دارد. در این مقاله، یک نسخه اصلاح شده از خوشه بندی مبتنی بر چگالی با استفاده از مفهوم "اشیاء هسته-قابل دسترسی با چگالی" و تخصیص مجدد اشیاء مرزی، برای حل این مشکل پیشنهاد شده است. با استفاده از این اصلاحات، الگوریتم خوشه بندی مبتنی بر چگالی اصلاح شده، نتایج خوشه بندی را بهبود می بخشد، به ویژه برای مجموعه داده هایی که شامل ساختارهای متراکم با خوشه های متصل هستند و پایداری بیشتر، خوشه بندی دقیق تر و عدم وابستگی به ترتیب پردازش را ارائه می دهد. در متن مقاله نیز به توانایی این

الگوریتم در تشخیص نقاط نویز در داده ها اشاره شده و این داده های نویزی توسط این الگوریتم به خوبی و به عنوان یک خوشه مجزا خوشه بندی می شوند. (تران^{۲۴} و همکاران، ۲۰۱۳)

روش تحقیق

در این پژوهش به جهت پیدا کردن داده های پرت، ما از یک روش ماشین لرنینگ به نام الگوریتم خوشه بندی مبتنی بر چگالی استفاده خواهیم کرد که این الگوریتم یک الگوریتم خوشه بندی است که بر اساس چگالی داده ها عمل می کند. این الگوریتم قادر به تشخیص نقاط پرت یا نویز است و می تواند خوشه هایی با اشکال مختلف را شناسایی کند، در حالی که الگوریتم های دیگری مانند K-Means تنها قادر به یافتن خوشه های کروی هستند و توانایی تشخیص داده های پرت را نیز ندارند. عملکرد این الگوریتم به شدت به انتخاب پارامترهای الگوریتم وابسته است. انتخاب نامناسب این پارامترها می تواند منجر به نتایج نامطلوب شود.

از مفاهیم کلیدی این الگوریتم، شعاع همسایگی^{۲۵}، فاصله ای است که برای تعیین همسایگی یک نقطه استفاده می شود، حداقل تعداد نقاط^{۲۶}، حداقل تعداد نقاط مورد نیاز برای تشکیل یک خوشه در همسایگی یک نقطه، نقطه هسته، نقطه ای است که حداقل تعداد نقاط در فاصله شعاع آن وجود دارد، نقطه مرزی، نقطه ای است که در فاصله شعاع یک نقطه هسته قرار دارد اما خودش نقطه هسته نیست، نقطه نویز، نقطه ای است که نه نقطه هسته و نه نقطه مرزی می باشد. در این بخش، تعدادی از تعاریف و ویژگی های الگوریتم خوشه بندی مبتنی بر چگالی مرور می شوند. با فرض مجموعه داده X حاوی N شیء، این الگوریتم یک الگوریتم خوشه بندی مبتنی بر چگالی است، بررسی بر اساس چگالی به این معنی است که الگوریتم خوشه بندی به جای تمرکز بر فاصله ی بین نقاط، به تراکم یا فشردگی نقاط در مناطق مختلف توجه می کند. در این روش، یک خوشه به عنوان منطقه ای با چگالی بالا تعریف می شود، یعنی جایی که تعداد زیادی نقطه در یک فضای محدود قرار دارند. نقاطی که در مناطق کم تراکم یا پراکنده قرار دارند، به عنوان نویز یا نقاط پرت در نظر گرفته می شوند. مزیت اصلی این رویکرد این است که می تواند خوشه هایی با اشکال مختلف و پیچیده را شناسایی کند، در حالی که روش های مبتنی بر فاصله (مثل روش خوشه بندی میانگین) معمولاً فقط خوشه های کروی شکل را به خوبی تشخیص می دهند.

این الگوریتم چگالی محلی را در همسایگی $x_i \in X$ با نماد $\text{density}(x_i)$ به صورت تعداد کل اشیاء در همسایگی آن فرموله می کند:

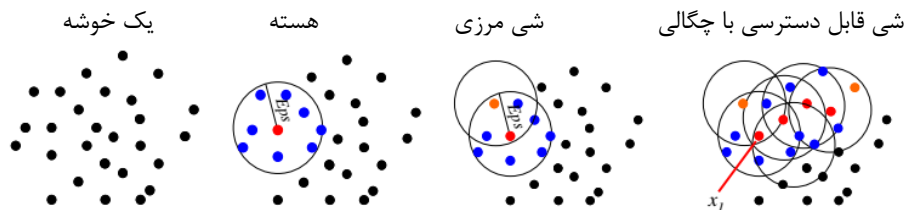
$$\text{density}(x_i) = \text{Count}(N_{\text{Eps}}(x_i))$$

استراتژی معاملاتی بر اساس خوشه بندی داده های قیمت... / شریفی، رستمی

که در آن $N_{Eps}(X_i)$ نشان دهنده اشیاء همسایه در همسایگی با شعاع Eps - اولین پارامتر این الگوریتم - است:

$$N_{Eps}(X_i) = \{x_j | \forall j, \text{distance}(x_j, x_i) < Eps\}$$

با در نظر گرفتن تعداد اشیاء در یک همسایگی، می توان سه نوع شیء را مشخص کرد، اشیاء هسته، اشیاء مرزی و نویزها. با توجه به حداقل تعداد اشیاء در همسایگی، که پارامتر دوم الگوریتم است، یک شیء X_{core} به عنوان یک شیء هسته تعریف می شود اگر $\text{density}(X_{core}) \geq \text{MinPts}$ (چگالی محلی آن بیشتر از حداقل تعداد نقاط است). یک شیء یک شیء مرزی، X_{border} است اگر $\text{density}(X_{border}) < \text{MinPts}$ و یک شیء هسته وجود داشته باشد به طوری که $X_{border} \in N_{Eps}(X_{core})$ ؛ یعنی شیء مرزی X_{border} به همسایگی X_{core} تعلق دارد و چگالی محلی آن کمتر از حداقل تعداد نقاط است. شکل ۱ یک نمای کلی از نحوه ی عملکرد این الگوریتم را نشان می دهد.



شکل ۱- جستجوی اشیاء با الگوریتم خوشه بندی مبتنی بر چگالی

یکی دیگر از تعاریف مهم مورد استفاده در خوشه بندی مربوط به اشیاء قابل دسترس از نظر چگالی است. دو شیء، X_1 و X_n ، قابل دسترس از نظر چگالی نامیده می شوند اگر یک زنجیره از اشیاء $X_1, X_2, \dots, X_n$ وجود داشته باشد که در آن $i \geq 1$ و $n \geq 2$ به طوری که برای همه $i \leq n$ ، X_i یک شیء هسته است، $\text{density}(X_i) \geq \text{MinPts}$ ، و X_{i+1} یک همسایه X_i است، $X_{i+1} \in N_{Eps}(X_i)$. به عنوان ویژگی ها، $X_1, \dots, X_{n-1}$ اشیاء هسته هستند و X_n یا یک شیء هسته است یا یک شیء مرزی. بر اساس تعاریف معرفی شده قبلی، یک خوشه از اشیاء به عنوان مجموعه ای از اشیاء تعریف می شود که از یک شیء هسته دلخواه (از یک خوشه) قابل دسترس از نظر چگالی هستند. آخرین دسته از اشیاء، اشیاء نویز هستند. یک شیء یک شیء نویز، X_{noise} ، است اگر در همسایگی با شعاع Eps داده شده، کمتر از حداقل تعداد نقاط شیء وجود داشته باشد که هیچ کدام شیء هسته نباشند $\text{density}(X_{noise}) < \text{MinPts}$. با استفاده از

الگوریتم، خوشه‌بندی اشیاء از هر خوشه C یک فرآیند دو مرحله‌ای است. در ابتدا، یک شیء هسته دلخواه از خوشه X_{c1} ، شناسایی می‌شود. سپس، تمام اشیاء قابل دسترس از نظر چگالی از شیء هسته X_{c1} بازیابی می‌شوند. در مرحله دوم، هر زنجیره از اشیاء شروع شده از X_{c1} به صورت بازگشتی شناسایی می‌شود.

به طور خلاصه و ساده میتون گفت این الگوریتم در ابتدا در مرحله مقداردهی اولیه، تمامی نقاط به عنوان "بازدید نشده" علامت‌گذاری می‌شوند، سپس یک نقطه "بازدید نشده" به صورت تصادفی انتخاب می‌شود و همسایگی در فاصله شعاع نقطه انتخاب شده بررسی می‌شود، اگر تعداد نقاط در همسایگی کمتر از حداقل تعداد نقاط باشد، نقطه به عنوان "نویز" علامت‌گذاری می‌شود، اگر تعداد نقاط در همسایگی بزرگتر یا مساوی حداقل تعداد نقاط باشد، یک خوشه جدید ایجاد می‌شود و نقطه انتخاب شده به عنوان "هسته" به این خوشه اضافه می‌شود، تمامی نقاط "بازدید نشده" در همسایگی شعاع نقطه هسته بررسی می‌شوند. اگر یک نقطه "بازدید نشده" در همسایگی یک نقطه هسته و حداقل تعداد نقاط نقطه در همسایگی شعاع خود را داشته باشد، آن نقطه نیز به عنوان "هسته" به خوشه اضافه می‌شود و همسایگی آن نیز به همین ترتیب بررسی می‌شود و گسترش خوشه ادامه می‌یابد. اگر یک نقطه "بازدید نشده" در همسایگی یک نقطه هسته کمتر از حداقل تعداد نقاط در شعاع خود داشته باشد، آن نقطه به عنوان "مرزی" به خوشه اضافه می‌شود. و سپس مراحل ۲ تا ۴ تا زمانی که تمامی نقاط "بازدید" شوند تکرار می‌شود.

به طور کلی از این الگوریتم برای تحلیل داده‌های مکانی و برای شناسایی خوشه‌های جغرافیایی، تشخیص تقلب و الگوهای غیرمعمول در داده‌ها که ممکن است نشان دهنده تقلب باشند، بخش‌بندی تصویر به مناطق مختلف بر اساس شباهت رنگ یا بافت استفاده شود و همچنین تشخیص نویز و نقاط پرت استفاده می‌شود.

یافته‌ها

در این پژوهش، برای انجام تحلیل‌ها و پیاده‌سازی روش‌های شناسایی داده‌های نویز، سهم‌های موجود در بازار بورس و فرابورس تهران مورد بررسی قرار گرفتند. هدف اصلی این انتخاب، اطمینان از کارایی روش‌های مورد استفاده در داده‌هایی با ویژگی‌های مختلف و فراهم آوردن بستری برای تعمیم‌پذیری نتایج بود.

در مسیر انجام تحلیل‌ها، پیش‌پردازش داده‌ها به عنوان گام اولیه مورد توجه قرار گرفت. در این مرحله:

استراتژی معاملاتی بر اساس خوشه بندی داده های قیمت... / شریفی، رستمی

- ویژگی‌های کلیدی با استفاده از فرآیند مهندسی ویژگی ها^{۲۷} استخراج شدند.
 - داده‌ها به منظور رفع تأثیر مقیاس‌های متفاوت استانداردسازی^{۲۸} شدند.
 - با بهره‌گیری از تحلیل همبستگی، متغیرهای وابسته شناسایی و حذف گردیدند.
 - برای تمرکز بر مهم‌ترین عوامل تأثیرگذار، فرآیند انتخاب ویژگی‌ها با روش جنگل تصادفی^{۲۹} اجرا شد.
 - نهایتاً داده‌های منتخب با روش‌های مناسب آماده‌سازی شدند تا برای اجرای الگوریتم‌های تحلیل مورد استفاده قرار گیرند.
- این اقدامات به عنوان بخش اساسی فرآیند آماده‌سازی داده، امکان اجرای دقیق‌تر و کارآمدتر الگوریتم‌های شناسایی نویز را فراهم آورد.
- در گام بعدی، الگوریتم خوشه‌بندی DBSCAN با هدف شناسایی نقاط نویز در داده‌ها اجرا شد. این الگوریتم با تنظیم پارامترهای شعاع همسایگی و تعداد حداقل نقاط در خوشه، داده‌ها را خوشه‌بندی کرد و نقاطی که خارج از خوشه‌ها قرار داشتند، به عنوان نویز شناسایی نمود. خروجی این مرحله شامل خوشه‌های اصلی و مجموعه‌ای از داده‌های نویز بود که رفتار غیرمعمولی در بازار نشان می‌دادند.
- برای استفاده از نتایج شناسایی نویز در طراحی یک استراتژی معاملاتی، داده‌های نویز استخراج شده از الگوریتم DBSCAN با سیگنال‌های تولیدشده توسط اندیکاتور MACD ترکیب شدند. این فرآیند به شکل زیر صورت گرفت:
- داده‌های نویز شناسایی شده توسط DBSCAN در داده‌های قیمتی علامت گذاری شدند.
 - اندیکاتور MACD، که یکی از ابزارهای تکنیکال رایج در تحلیل بازار است، برای شناسایی نقاط ورود و خروج بهینه به معاملات مورد استفاده قرار گرفت.
 - نقاطی که هم‌زمان با وجود نویز و تغییرات در سیگنال MACD هم‌پوشانی داشتند، به عنوان نقاط ورود و خروج برای معاملات علامت‌گذاری شدند.
- در ادامه به شرح و نمایش نتایج هر مرحله خواهیم پرداخت.
۱. آماده‌سازی داده‌ها و مهندسی ویژگی‌ها
- در ابتدا، داده‌های اولیه از بازار مالی جمع‌آوری شدند که شامل متغیرهایی نظیر حجم معاملات، قیمت، و سایر ویژگی‌های مرتبط بودند. برای اطمینان از کیفیت داده‌ها و وارد کردن ویژگی‌هایی با بار اطلاعاتی مناسب برای آموزش الگوریتم کار ساخت و مهندسی ویژگی‌های جدید انجام شد. ویژگی‌های های ساخته شده از ترکیب اطلاعات مربوط به قیمت و حجم معاملات ایجاد شده است که شامل موارد زیر است:
- نسبت حجم معاملات تعدیل شده به فاصله بین قیمت باز شدن و بسته شدن

$$\text{Ratio } ۱ = \frac{\text{Adj Volume}}{|\text{Open}-\text{Close}|}$$

Adj Volume حجم معاملات تعدیل شده با افزایش سرمایه و توزیع سود نقدی است.

- نسبت حجم معاملات تعدیل شده به فاصله بین بیشترین و کمترین قیمت در روز

$$\text{Ratio } ۲ = \frac{\text{Adj Volume}}{|\text{High}-\text{Low}|}$$

- بازده روزانه مثبت یا منفی باشد

$$\text{Positive Return} = \begin{cases} ۱ & \text{if Return} > ۰ \\ ۰ & \text{else} \end{cases}$$

- فاصله بین بیشترین و کمترین قیمت در روز

$$\text{Range} = \text{High} - \text{Low}$$

- میانگین قیمت روز

$$\text{Average Price} = \frac{\text{Open} + \text{Close} + \text{High} + \text{Low}}{۴}$$

- نسبت نوسان قیمت روزانه به قیمت بسته شدن

$$\text{Volatility Ratio} = \frac{\text{Range}}{\text{Close}}$$

- نسبت حجم تعدیل شده به تفاوت قیمت باز و بسته شدن، وزن دار با قیمت باز شدن

$$\text{Open Adj Volume Ratio} = \frac{\text{Open} \times \text{Adj Volume}}{|\text{Open} - \text{Close}|}$$

- لگاریتم ویژگی Open Adj Volume_Ratio

$$\text{Log (Open Adj Volume Ratio)}$$

- نسبت حجم تعدیل شده وزن دار با میانگین قیمت باز و بسته شدن

$$\text{Open Adj Volume Ratio Mean} = \frac{\text{Open} \times \text{Adj Volume}}{\text{Open} + \text{Close}}$$

- نسبت حجم تعدیل شده به میانگین حجم ۷ روز گذشته

$$\text{Volume Ratio} = \frac{\text{Adj Volume}}{\text{Adj Volume}(\text{Rolling mean, window} = ۷)}$$

- نسبت دامنه روزانه قیمت به میانگین دامنه ۷ روز گذشته

استراتژی معاملاتی بر اساس خوشته بندی داده های قیمت.../ شریفی، رستمی

$$\text{Range Ratio} = \frac{\text{Range}}{\text{Range(Rolling mean, window = ۷)}}$$

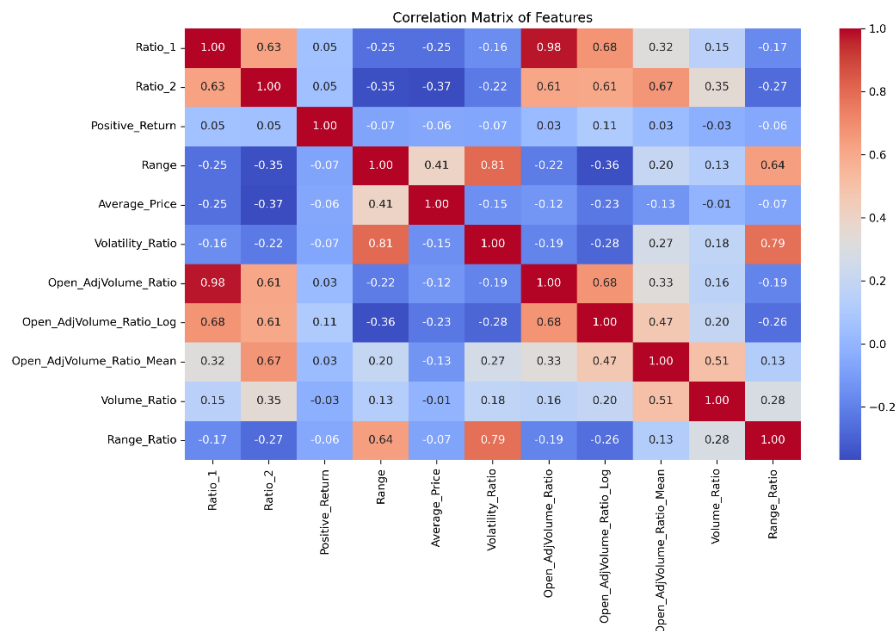
۲. استانداردسازی داده‌ها برای کاهش تأثیر مقیاس‌های مختلف داده‌های ورودی

۳. شناسایی و حذف متغیرهای همبسته با استفاده از تحلیل همبستگی^{۳۰}

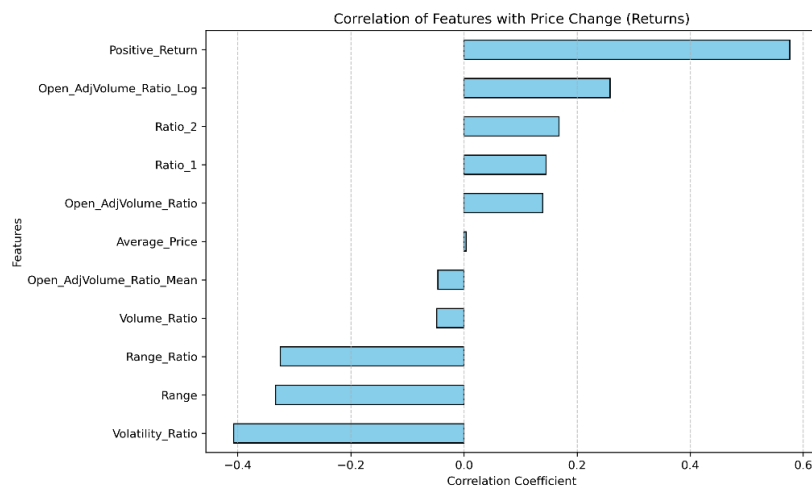
در این قسمت پس از ساخت فیچر ها از داده های قیمت و حجم، برای هر سهم مورد نظر ماتریس همبستگی بین ویژگی ها محاسبه شد، ماتریس همبستگی ارائه شده، روابط میان فیچرهای مختلف را بر اساس ضریب همبستگی پیرسون نشان می‌دهد. این تحلیل به ما کمک می‌کند تا بفهمیم کدام فیچرها ممکن است اطلاعات مشابه ارائه دهند (همبستگی بالا) و کدام یک به طور مستقل می‌توانند در مدل استفاده شوند (همبستگی پایین).

ماتریس همبستگی برای نماد فارس از بازار بوس تهران در شکل ۲ قابل مشاهده است. ماتریس همبستگی نشان می‌دهد که Ratio Volatility و Range Ratio دارای همبستگی مثبت قوی (۰٫۷۰) هستند، در حالی که Average Price با ۲ Ratio همبستگی منفی قابل توجهی (-۰٫۵۲) دارد. ویژگی‌های با همبستگی بالا مانند ۱ Ratio و Open Adj Volume Ratio ممکن است اطلاعات مشابهی ارائه دهند و قابل ساده‌سازی باشند. در مقابل، متغیرهای مستقل‌تر مانند Volume Ratio می‌توانند ارزش افزوده بیشتری در مدل‌سازی داشته باشند. این نتایج برای بهینه‌سازی مدل و کاهش ابعاد داده‌ها مفید است.

همچنین در ادامه همبستگی تمام فیچرهای ساخته شده با بازدهی نیز محاسبه شده که در شکل ۳ آورده شده است. در این تحلیل، همبستگی ویژگی‌های مختلف با تغییرات قیمت (بازدهی) بررسی شده است. Return Positive بالاترین همبستگی مثبت را نشان می‌دهد، که بیانگر تأثیر قوی آن بر بازدهی است. ویژگی‌هایی مانند Volatility Ratio و Range Ratio همبستگی منفی قابل توجهی دارند، نشان‌دهنده این که افزایش این متغیرها ممکن است کاهش بازدهی را به دنبال داشته باشد. در مقابل، متغیرهایی مانند Average Price و Volume Ratio همبستگی ضعیفی داشته و تأثیر کمتری بر تغییرات قیمت دارند. این نتایج نشان می‌دهد که برای مدل‌سازی دقیق، تمرکز بر ویژگی‌های با همبستگی قوی (مثبت یا منفی) اولویت دارد، در حالی که ویژگی‌های کم‌اثر ممکن است حذف یا ساده‌سازی شوند.



شکل ۲- ماتریس همبستگی ویژگی‌ها

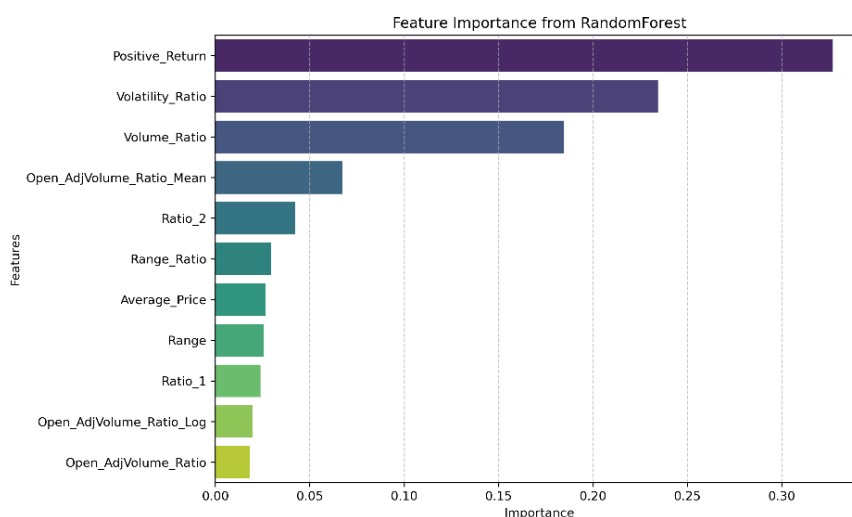


شکل ۳- همبستگی ویژگی‌ها با تغییرات قیمت

۴. انتخاب ویژگی‌های کلیدی با استفاده از الگوریتم جنگل تصادفی^{۳۱}

استراتژی معاملاتی بر اساس خوشه بندی داده های قیمت.../ شریفی، رستمی

نمودار اهمیت ویژگی ها^{۳۲} میزان تأثیر هر ویژگی را بر پیش بینی مدل نشان می دهد. این نمودار بر اساس مدل جنگل تصادفی محاسبه شده است و مشخص می کند کدام ویژگی ها بیشترین نقش را در پیش بینی هدف ایفا می کنند. ویژگی های با اهمیت بیشتر سهم بیشتری در تصمیم گیری مدل دارند. در این نمودار در که در مورد سهم فارس از بازار بورس تهران صورت گرفته و در شکل ۴ نشان داده شده، Positive Return, Volatility Ratio و Range Ratio از مهمترین ویژگی ها هستند و ویژگی هایی مانند Range و Average Price در انتهای لیست قرار گرفته اند و نقش کمتری در مدل دارند. این اطلاعات می تواند برای بهینه سازی مدل و کاهش پیچیدگی داده ها مفید باشد، به این صورت که ویژگی های کم اهمیت حذف شوند و تمرکز بر ویژگی های مهم تر قرار گیرد.

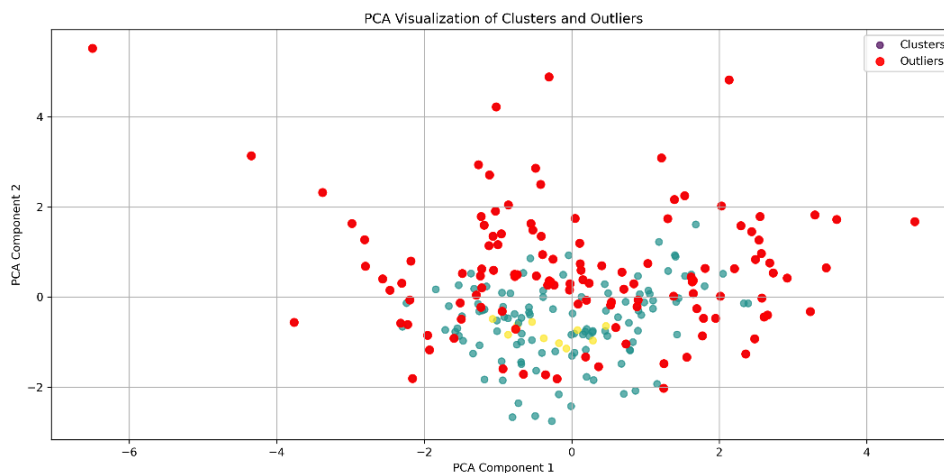


شکل ۴- بررسی اهمیت ویژگی ها با الگوریتم جنگل تصادفی

۵. کاهش ابعاد با PCA^{۳۳}

تحلیل PCA یا "تحلیل مؤلفه های اصلی" برای کاهش ابعاد داده ها و ساده سازی نمایش آن ها نشان می دهد که داده ها عمدتاً در نزدیکی مرکز متمرکز هستند، اما برخی نقاط به صورت پراکنده در فواصل دورتر قرار گرفته اند که می توانند نشان دهنده نویز یا خوشه های کوچک باشند. این کاهش ابعاد، خوشه بندی داده ها با الگوریتم DBSCAN را ساده تر کرده و امکان شناسایی الگوهای پنهان را فراهم می کند. خروجی مربوط به این بخش در شکل ۵ قابل مشاهده است. همچنین این تحلیل نشان می دهد که داده های مالی پس از کاهش ابعاد، الگوهای قابل تفکیکی را نشان می دهند. توزیع داده ها در این فضا بیانگر پتانسیل

مناسب الگوریتم DBSCAN برای خوشه‌بندی آن‌ها است. پراکندگی نقاط به صورت خوشه‌های متراکم و نقاط دورافتاده (نویز) نشان می‌دهد که DBSCAN، با توانایی در شناسایی خوشه‌های غیرخطی و نقاط نویز، می‌تواند به طور موثری در تحلیل و خوشه‌بندی داده‌های مالی به کار گرفته شود. این ترکیب ابزارها امکان استخراج الگوهای پنهان از داده‌های پیچیده مالی را فراهم می‌کند.



شکل ۵- بصری‌سازی تحلیل مؤلفه‌های اصلی برای خوشه‌ها و داده‌های پرت

تمام مراحل مربوط به ساخت و انتخاب ویژگی‌های مهم برای نمادهای دیگر نیز صورت گرفته و برای آموزش الگوریتم مورد استفاده قرار گرفته است. و پس از اتمام مراحل بالا ویژگی‌های زیر که برای تمام نمادهای مورد بررسی بهترین اثرگذاری را داشته‌اند، برای ورود به الگوریتم انتخاب شده است. Positive Return, Volatility Ratio, Volume Ratio, Range Ratio, open AdjVolume, Ratio log, Ratio^۲.

۶. شناسایی خوشه‌ها و نقاط نویز با روش DBSCAN

همان‌طور که گفته شد در این پژوهش از مدل DBSCAN برای شناسایی خوشه‌های داده و نقاط نویز مورد استفاده قرار گرفت. داده‌های ورودی به الگوریتم، ویژگی‌های مهم استخراج شده از قسمت مهندسی ویژگی‌ها می‌باشد و نتایج مربوط به بخش خوشه‌بندی پس از استخراج به شکل یک استراتژی معاملاتی مورد استفاده قرار می‌گیرد.

داده‌های پرت در واقع به عنوان نقاطی با ویژگی‌های متفاوت از سایر نقاط انتخاب شده است و یک علامت مهم برای ورود یا خروج از معامله در یک استراتژی قیمتی می‌تواند باشد. برای تعیین داده‌های پرت توسط این الگوریتم ما داده‌های قیمتی از تاریخ ۱۳۹۸/۰۹/۱۵ تا ۱۴۰۲/۰۹/۱۴ را به الگوریتم برای

استراتژی معاملاتی بر اساس خوشه بندی داده های قیمت... / شریفی، رستمی

آموزش داده‌ایم و با این داده‌ها پارامترهای بهینه‌الگوریتم را با روش جستجوی شبکه‌ای^{۳۴} یافته و شبکه را برای تست داده‌های آینده و دیده نشده^{۳۵} ذخیره می‌کنیم. سپس برای تست شبکه داده‌ها از تاریخ ۱۴۰۲/۰۹/۱۵ تا ۱۴۰۳/۰۹/۱۵ را به صورت روز به روز به شبکه‌ی آموزش دیده وارد کرده و الگوریتم، خوشه‌ی داده جدید را مشخص می‌کند. اگر داده جدید به عنوان یک داده نویز خوشه بندی می‌شد، نقطه ورود یا خروج از معامله مشخص می‌شد و سپس برای تعیین ورود یا خروج از معامله از سیگنال ورود و خروج^{۳۶} MACD استفاده کرده و اگر در هر نقطه نویز، مکدی در سیگنال ورود به معامله داشت یعنی زمانی که خط MACD از زیر خط سیگنال عبور کرده بود، دستور ورود به خرید صادر و اگر در زمان ایجاد شدن داده نویز، خط MACD از بالا خط سیگنال را قطع کرده بود، دستور خروج از معامله ثبت می‌گردد. پس از اتمام بررسی کل داده‌های شبکه تست، بین نقاط ورود و خروج بازدهی محاسبه شده تا میزان موفقیت استراتژی بررسی گردد.

در این مطالعه، علاوه بر استفاده از روش خوشه‌بندی DBSCAN برای تشخیص داده‌های نویز، از یک روش آماری نیز بهره گرفته شده است. در این روش، بر اساس یکی از ویژگی‌های کلیدی داده‌ها که از توزیع نرمال پیروی می‌کند، نقاطی که خارج از یک انحراف معیار از میانگین قرار داشتند به عنوان داده‌های پرت در نظر گرفته شدند. هدف از این بررسی آماری، مقایسه نتایج آن با روش DBSCAN برای ارزیابی دقت و کارایی هر دو روش در شناسایی نویزها بود. به این روش در طول پژوهش، تحلیل انحراف معیار گاوسی^{۳۷} گفته خواهد شد.

علاوه بر این، عملکرد هر دو روش DBSCAN و روش آماری با استراتژی خرید و نگهداری^{۳۸} نیز مقایسه شده است تا مشخص شود که کدام روش در بهبود بازدهی سرمایه‌گذاری و مدیریت ریسک مؤثرتر عمل می‌کند. این مقایسه به ما امکان داد تا دید جامع‌تری از کاربرد این تکنیک‌ها در تحلیل داده‌های مالی به دست آوریم.

در جدول ۱ بازدهی به دست آمده از سه استراتژی DBSCAN و تحلیل انحراف معیار گاوسی و خرید و نگهداری برای ۱۰ سهم پرتراکنش بازار بورس تهران، و در جدول ۲ بازدهی استراتژی‌های مذکور برای ۱۰ نماد پرتراکنش بازار فرابورس آورده شده است. همچنین در

جدول ۳ نمادهایی با بیشترین بازدهی نسبت به خرید و نگهداری از روش DBSCAN آورده شده، این نمادها اختلاف بازدهی بیشتر از ۵۰ درصد با استراتژی DBSCAN نسبت به خرید و نگهداری دارند. و در

جدول ۴ نمادهایی با بیشترین بازدهی از روش انحراف معیار گاوسی آورده شده است. این نمادها اختلاف بازدهی بیشتر از ۴۰ درصد با استراتژی انحراف معیار گاوسی نسبت به خرید و نگهداری دارند.

برای بررسی عملکرد کلی استراتژی ها در ابتدا به بررسی جدول ۱ می پردازیم، در مورد استراتژی DBSCAN و MACD، این استراتژی بازدهی بالایی در اکثر نمادها ارائه کرده است. به خصوص در نمادهای:

وتجارت: بازدهی ۷۹.۷۸٪، با اختلاف بازدهی ۶۵.۸۸٪ نسبت به خرید و نگهداری.
فجر: بازدهی ۵۱.۴۴٪، با اختلاف بازدهی ۵۲.۵۲٪ نسبت به خرید و نگهداری.
های وب: بازدهی ۱۹.۳۳٪، با اختلاف بازدهی ۴۷.۳۶٪ نسبت به خرید و نگهداری.
شستا: بازدهی ۵۰.۲۰٪، با اختلاف بازدهی ۳۳.۴۷٪ نسبت به خرید و نگهداری.
و نکته قابل توجه این است که اختلاف مثبت بازدهی با خرید و نگهداری در بیشتر نمادها مشهود است، که نشان دهنده کارایی این استراتژی در شناسایی فرصت های سودده است.

در مورد استراتژی GSDM و MACD:

بازدهی این استراتژی در برخی نمادها مثبت و در برخی دیگر منفی بوده است:
در نمادهایی مانند فجر ۲۳.۵۸٪ و های وب ۱۷.۶۶٪ عملکرد قابل قبولی دارد.
اما در نمادهایی مانند شاهد بازدهی منفی ۲.۱۹٪- ثبت شده است.
اختلاف بازدهی این استراتژی نسبت به خرید و نگهداری مثبت بوده است، اما به طور میانگین کمتر از DBSCAN + MACD عمل کرده است.

در مورد استراتژی خرید و نگهداری، این روش در اکثر نمادها بازدهی منفی یا بسیار پایین داشته است:
برای مثال: های وب ۲۸.۰۳٪- و شاهد ۲۴.۶۳٪-

ضعف این استراتژی خصوصا در شرایط نوسانی و رنج بازار کاملا مشهود است.
بررسی جدول ۲ در مورد نمادهای فرابورس، در مورد استراتژی DBSCAN و MACD، در این دسته از نمادها در بازار فرابورس نیز عملکرد نسبتا خوبی داشته و در اکثر نمادها بازدهی مثبت یا بالاتر از سایر روش ها ارائه کرده است:

فرابورس: بازدهی ۳۸.۲۸٪، اختلاف بازدهی مثبت ۵۴.۳۳٪ نسبت به خرید و نگهداری.
عالیس: بازدهی ۲۹.۴۷٪، اختلاف بازدهی ۳۳.۸۸٪.
وسپهر: بازدهی ۲۶.۹۱٪، با اختلاف ۲۵.۸۷٪.
و در این جدول نیز این استراتژی در بیشتر نمادها اختلاف مثبت قابل توجهی با خرید و نگهداری نشان داده است.

استراتژی GSDM و MACD در نمادهای پرتراکنش فرابورس

عملکرد این استراتژی در برخی نمادها مناسب بوده اما در برخی دیگر ضعیف عمل کرده است:
فرابورس: بازدهی ۱۲.۰۳٪، اختلاف بازدهی ۲۸.۰۹٪
عالیس: بازدهی ۱۳.۶۹٪، اختلاف بازدهی ۱۸.۰۹٪

استراتژی معاملاتی بر اساس خوشه بندی داده های قیمت.../ شریفی، رستمی

وسپهر :بازدهی ۳۸.۶۵٪، که در این نماد از DBSCAN بهتر عمل کرده است.

بررسی

جدول ۳ در مورد عملکرد کلی استراتژی DBSCAN و MACD و نمادهای با بیشترین اختلاف بازدهی این استراتژی و خرید و نگه داری را نشان می دهد.

این استراتژی در اکثر نمادهای ارائه شده بازدهی بسیار بالایی داشته است.

دکوتر :بازدهی فوق العاده ۱۲۲.۵۸٪، با اختلاف بازدهی ۹۳.۸۶٪ نسبت به خرید و نگه داری.

خوساز :بازدهی ۸۸.۶۳٪، با اختلاف بازدهی ۸۷.۱۵٪

وآتوس :بازدهی ۴۳.۴۹٪، با اختلاف بازدهی ۱۰۵.۴۸٪

مشخص است این استراتژی در نمادهایی با پتانسیل بازدهی بالا، به خوبی عمل کرده و اختلاف مثبت قابل توجهی نسبت به سایر روش ها دارد.

استراتژی GSDM و MACD :

بازدهی این استراتژی در کل کمتر از DBSCAN بوده اما در برخی نمادها همچنان عملکرد قابل قبولی داشته است:

دکوتر :بازدهی ۴۴.۴۹٪ (اختلاف مثبت ۱۵.۷۶٪ نسبت به خرید و نگه داری).

خوساز :بازدهی ۴۶.۳۸٪، که در مقایسه با DBSCAN بازدهی کمتری دارد.

وآتوس :بازدهی ۲.۳۷٪، که نشان دهنده ضعف این روش در این نماد است.

خرید و نگه داری:

این روش در بسیاری از نمادها بازدهی منفی یا بسیار پایین داشته است:

وآتوس :بازدهی منفی ۶۱.۹۹٪-

دکوتر :بازدهی ۲۸.۷۳٪، که در مقایسه با DBSCAN بسیار کمتر است.

خوساز :بازدهی ۱.۴۸٪، که نشان دهنده ناکارآمدی این روش در این نمادها است.

و بررسی جدول ۴ که نشان دهنده نمادهایی بیشترین اختلاف بازدهی روش انحراف معیار گاوسی با خرید و نگهداری است، به طور کلی این استراتژی عملکردی پایین تر از روش DBSCAN داشته ولی در برخی نمادها از DBSCAN بهتر عمل کرده است. و در بیشتر نمادها عملکرد بهتری از خرید و نگهداری از خود نشان داده است.

عملکرد کلی استراتژی GSDM و MACD :

این استراتژی در این جدول بالاترین بازدهی را نسبت به سایر روش ها در بسیاری از نمادها ارائه کرده است:

ثنوسا :بازدهی ۸۶.۱۵٪، که نسبت به خرید و نگه داری (۸.۴۴٪) اختلاف مثبت ۷۷.۷۲٪ داشته است.

حفارس :بازدهی ۷۰.۴۶٪، با اختلاف مثبت ۷۴.۱۰٪ نسبت به خرید و نگه داری.

نکته کلیدی: این استراتژی در نمادهای دارای پتانسیل رشد بالا عملکرد بسیار مطلوبی داشته است. عملکرد کلی استراتژی DBSCAN و MACD: در این جدول، بازدهی این استراتژی در برخی نمادها نزدیک به GSDM بوده اما همچنان کمتر از آن است:

ثنوسا: بازدهی ۶۷.۹۷٪، اختلاف بازدهی ۵۹.۵۳٪ نسبت به خرید و نگهداری.
حفارس: بازدهی ۵۳.۵۹٪، با اختلاف بازدهی ۵۷.۲۳٪
و در مورد عملکرد کلی استراتژی خرید و نگهداری، این روش در این نمادها بازدهی پایین یا حتی منفی داشته است:

ثنوسا: بازدهی ۸.۴۴٪، که در مقایسه با $GSDM + MACD$ بسیار پایین تر است.
حفارس: بازدهی منفی ۳.۶۴٪، که نشان دهنده ناکارآمدی این روش در شرایط خاص بازار است.
در ادامه در نمودار ۱ و نمودار ۲ مقادیر جداول ۱ و ۳ برای مقایسه بهتر آورده شده.
همچنین برای نمایش بصری نحوه عملکرد این استراتژی، در شکل ۶ و شکل ۷ نمودار قیمتی نماد ذوب با نقاط داده پرت مشخص شده از دو استراتژی آورده شده است. شبکه آموزش و تست در این نمودار قابل مشاهده است و نمودار مربوط به MACD نیز در زیر نمودار قیمت قابل مشاهده است. همانطور که در شکل مشخص شده، نقاط سبز رنگ، داده های پرت هستند که با MACD علامت ورود به معامله را مشخص میکند و در نقاط قرمز رنگ، داده های پرت با سیگنال خروج MACD، علامت خروج از استراتژی را صادر میکنند.

استراتژی معاملاتی بر اساس خوشه بندی داده های قیمت.../ شریفی، رستمی

جدول ۱ - مقایسه بازدهی سه استراتژی در نمادهای پرتراکنش بورس تهران (اعداد به درصد)

نماد	DBSCAN + MACD	GSDM + MACD	خرید و نگه داری	اختلاف بازدهی و DBSCAN خرید و نگهداری	اختلاف بازدهی و GSDM خرید و نگه داری
وتجارت	۷۹.۷۸	۶۷.۲۷	۱۳.۸۹	۶۵.۸۸	۵۳.۳۷
ثشاهد	۲۹.۶۴	-۲.۱۹	-۲۴.۶۳	۵۴.۲۷	۲۲.۴۳
فجر	۵۱.۴۴	۲۳.۵۸	-۱.۰۷	۵۲.۵۲	۲۴.۶۶
های وب	۱۹.۳۳	۱۷.۶۶	-۲۸.۰۳	۴۷.۳۶	۴۵.۶۹
ذوب	۵۴.۸۹	۴۵.۴۹	۸.۵۶	۴۶.۳۲	۳۶.۹۲۱
برکت	۳.۱۹	-۱۲.۴۵	-۴۰.۷۰	۴۳.۸۹	۲۸.۲۴
خگستر	۳۳.۲۵	۱۷.۳۵	-۶.۴۲۸	۳۹.۶۷	۲۳.۷۸
اخابر	۲۱.۷۲	۱۱.۳۱	-۱۷.۶۵	۳۹.۳۸	۲۸.۹۷
شستا	۵۰.۲۰	۶۰.۳۱	۱۶.۷۲	۳۳.۴۷	۴۳.۵۸
شپنا	۱۸.۰۳	-۰.۶۲	-۱۲.۹۴	۳۰.۹۸	۱۲.۳۲

جدول ۲ - مقایسه بازدهی سه استراتژی در نمادهای پرتراکنش فرابورس (اعداد به درصد)

نماد	DBSCAN + MACD	GSDM + MACD	خرید و نگه داری	اختلاف بازدهی و DBSCAN خرید و نگهداری	اختلاف بازدهی و GSDM خرید و نگه داری
فرابورس	۳۸.۲۸	۱۲.۰۳	-۱۶.۰۵	۵۴.۳۳	۲۸.۰۹
حآفرین	۱۰.۰۷۳	-۲۵.۶۰	-۳۳.۳۱	۴۳.۳۹	۷.۷۱
عالیس	۲۹.۴۷	۱۳.۶۹	-۴.۴۰	۳۳.۸۸	۱۸.۰۹
وسپهر	۲۶.۹۱	۳۸.۶۵	۱.۰۳	۲۵.۸۷	۳۷.۶۱
سرچشمه	۲۷.۴۷	۱۷.۵۹	۱.۷۴	۲۵.۷۳	۱۵.۸۴
گدنا	-۴.۰۴	۱۹.۸۰	-۲۶.۰۹	۲۲.۰۵	۴۵.۹۰
خاور	-۲۹.۵۴	-۴۴.۲۲	-۵۰.۶۴	۲۱.۰۹	۶.۴۱
سمگا	۴۵.۸۱	۱۳.۸۴	۳۳.۲۶	۱۲.۵۴	-۱۹.۴۲
فن افزار	۱۹.۹۳	۳۴.۵۱	۱۴.۱۸	۵.۷۴۱	۲۰.۳۳
داوه	۱۷.۵۴	-۳.۱۷	۳۵.۵۴	-۱۸.۰۰	-۳۸.۷۱

جدول ۳ - نمادهایی با بیشترین اختلاف بازدهی استراتژی DBSCAN و خرید و نگهداری (اعداد به درصد)

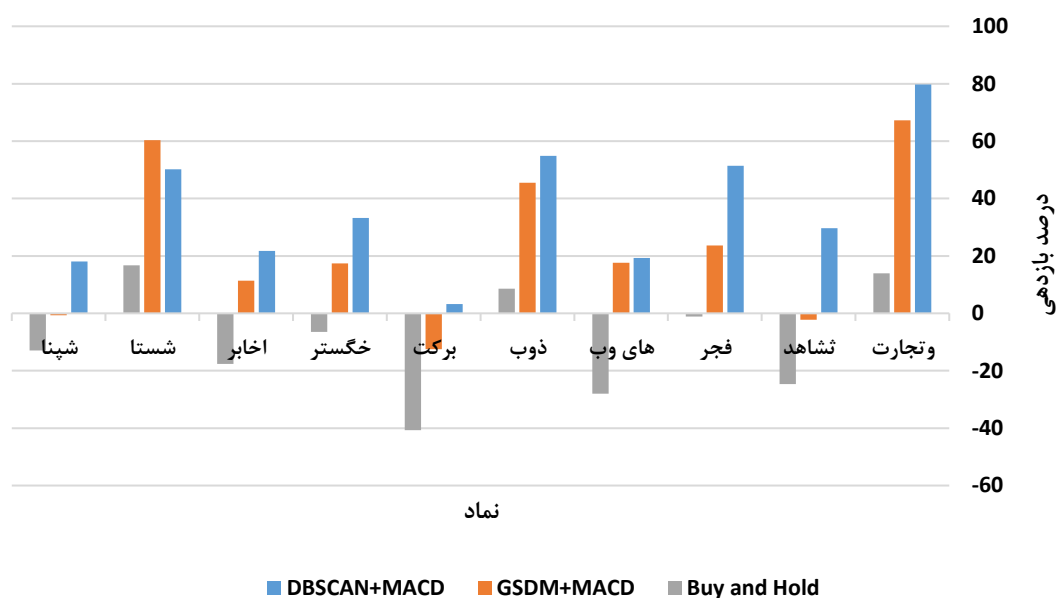
نماد	DBSCA N + MACD	GSDM + MACD	خرید و نگه داری	اختلاف بازدهی و DBSCAN خرید و نگهداری	اختلاف بازدهی و خرید و نگه داری GSDM
وآتوس	۴۳.۴۹	۲.۳۷	-۶۱.۹۹	۱۰۵.۴۸	۶۴.۳۶
دکوتر	۱۲۲.۵۸	۴۴.۴۹	۲۸.۷۳	۹۳.۸۶	۱۵.۷۶
خوساز	۸۸.۶۳	۴۶.۳۸	۱.۴۸	۸۷.۱۵	۴۴.۹۰
سدبیر	۱۴.۷۸	۷.۰۰	-۵۳.۸۸	۶۸.۶۶	۶۰.۸۸
خدیزل	۳۸.۲۲	-۲.۰۷	-۳۰.۱۶	۶۸.۳۹	۲۸.۰۹
وتجارت	۷۹.۷۸	۶۷.۲۸	۱۳.۹۰	۶۵.۸۹	۵۳.۳۸
غشهداب	۳۱.۹۶	۱۵.۵۷	-۳۳.۱۸	۶۵.۱۴	۴۸.۷۵
کیمیا	۴۲.۳۳	۱۵.۲۱	-۲۱.۵۷	۶۳.۹۰	۳۶.۷۹
کماسه	۴۲.۱۷	۶.۰۱	-۲۰.۹۳	۶۳.۱۰	۲۶.۹۴
ولشرق	۲۴.۲۱	-۲.۱۲	-۳۸.۸۳	۶۳.۰۴	۳۶.۷۱
ثنوسا	۶۷.۹۷	۸۶.۱۵	۸.۴۴	۵۹.۵۳	۷۷.۷۲
چکاپا	۱۳.۸۳	-۸.۲۸	-۴۴.۹۶	۵۸.۷۹	۳۶.۶۸
فاسمین	۳۶.۳۳	۹.۲۳	-۲۲.۱۸	۵۸.۵۲	۳۱.۴۱
توسن	-۲.۳۶	-۳۶.۴۶	-۶۰.۸۶	۵۸.۵۰	۲۴.۳۹
حفارس	۵۳.۵۹	۷۰.۴۶	-۳.۶۴	۵۷.۲۳	۷۴.۱۰
دارو	۲۳.۰۱	-۵.۷۱	-۳۴.۱۴	۵۷.۱۵	۲۸.۴۳
فاذر	۲۵.۸۸	۱۷.۳۶	-۳۱.۲۴	۵۷.۱۱	۴۸.۶۰
سکرما	۹۵.۵۴	۴۶.۷۸	۴۰.۵۸	۵۴.۹۷	۶.۲۰
سجام	۱۳.۰۳	-۹.۳۲	-۴۱.۷۰	۵۴.۷۴	۳۲.۳۸
شلعاب	۳۸.۶۴	۳۳.۰۹	-۱۵.۷۴	۵۴.۳۸	۴۸.۸۳
فراپورس	۳۸.۲۹	۱۲.۰۴	-۱۶.۰۵	۵۴.۳۴	۲۸.۰۹
ثشاهد	۲۹.۶۴	-۲.۲۰	-۲۴.۶۳	۵۴.۲۸	۲۲.۴۴
ولساپا	۲۵.۸۴	-۹.۸۸	-۲۸.۳۰	۵۴.۱۴	۱۸.۴۲
شوینده	۴۸.۵۳	۲۴.۳۸	-۴.۸۶	۵۳.۳۸	۲۹.۲۴
شفا	۱۲.۸۴	-۵.۶۰	-۳۹.۸۳	۵۲.۶۷	۳۴.۲۲

استراتژی معاملاتی بر اساس خوشته بندی داده های قیمت.../ شریفی، رستمی

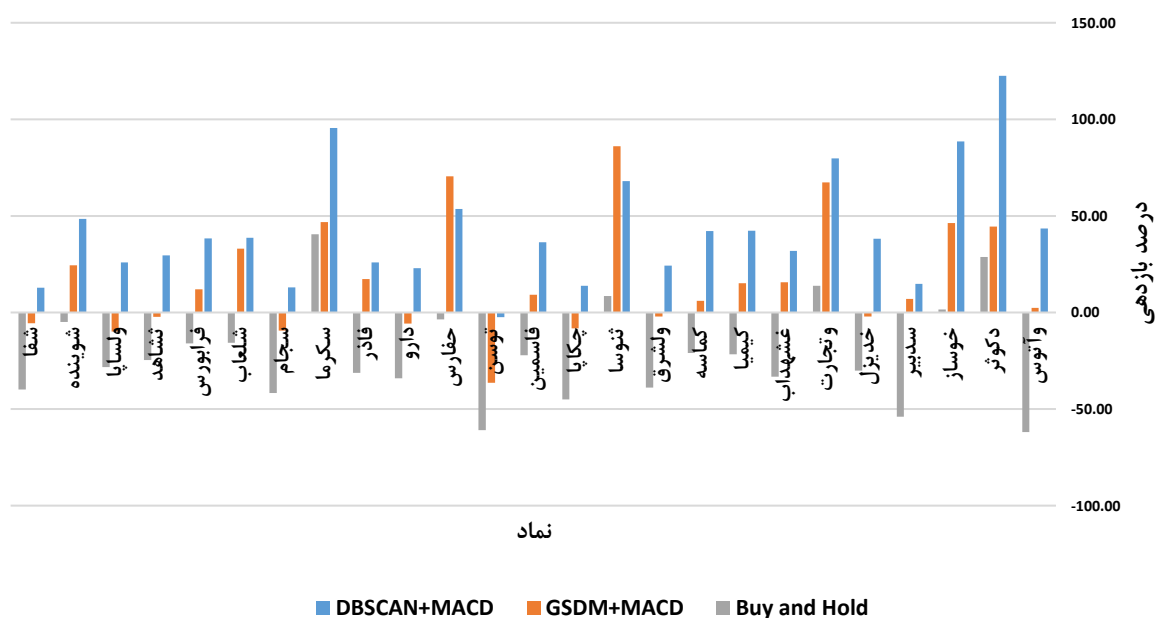
جدول ۴- نمادهایی با بیشترین اختلاف بازدهی استراتژی GSDM و خرید و نگهداری (اعداد به درصد)

نماد	DBSCAN + MACD	GSDM + MACD	خرید و نگه داری	اختلاف بازدهی و DBSCAN خرید و نگهداری	اختلاف بازدهی GSDM و خرید و نگه داری
ثنوسا	۶۷.۹۷	۸۶.۱۵	۸.۴۴	۵۹.۵۳	۷۷.۷۲
حفارس	۵۳.۵۹	۷۰.۴۶	-۳.۶۴	۵۷.۲۳	۷۴.۱۰
وآتوس	۴۳.۴۹	۲.۳۷	-۶۱.۹۹	۱۰۵.۴۸	۶۴.۳۶
کسرا	۳۹.۴۳	۵۱.۹۴	-۱۰.۲۹	۴۹.۷۲	۶۲.۲۳
سدبیر	۱۴.۷۸	۷.۰۰	-۵۳.۸۸	۶۸.۶۶	۶۰.۸۸
ثامن	۱۱.۳۹	۱۶۰.۳۷	۱۰۳.۲۹	-۹۱.۹۰	۵۷.۰۸
وتجارت	۷۹.۷۸	۶۷.۲۸	۱۳.۹۰	۶۵.۸۹	۵۳.۳۸
خبهمن	۴۳.۳۷	۴۳.۶۱	-۸.۵۶	۵۱.۹۲	۵۲.۱۶
وتوس	۰.۰۰	۵۲.۳۲	۰.۸۳	-۰.۸۳	۵۱.۴۹
ولملت	-۱۳.۶۸	۲.۹۱	-۴۸.۴۷	۳۴.۷۹	۵۱.۳۸
ولصنم	۲۶.۶۶	۳۸.۳۶	-۱۲.۳۴	۳۹.۰۰	۵۰.۶۹
فولای	۵۷.۷۷	۶۲.۵۰	۱۲.۳۵	۴۵.۴۳	۵۰.۱۶
بوس	-۸.۴۰	۴۸.۰۷	-۱.۶۵	-۶.۷۵	۴۹.۷۲
شلعاب	۳۸.۶۴	۳۳.۰۹	-۱۵.۷۴	۵۴.۳۸	۴۸.۸۳
غشهداب	۳۱.۹۶	۱۵.۵۷	-۳۳.۱۸	۶۵.۱۴	۴۸.۷۵
فاذر	۲۵.۸۸	۱۷.۳۶	-۳۱.۲۴	۵۷.۱۱	۴۸.۶۰
واحیا	۲۴.۴۷	۴۴.۲۱	-۴.۲۴	۲۸.۷۱	۴۸.۴۵
فسرب	۸.۴۴	۴۰.۶۷	-۷.۷۸	۱۶.۲۲	۴۸.۴۵
فرآور	۱۴.۰۴	۱۴.۳۲	-۳۳.۹۰	۴۷.۹۴	۴۸.۲۲
وسینا	۵۵.۸۹	۵۱.۶۹	۴.۴۷	۵۱.۴۲	۴۷.۲۲
فکمند	-۷.۵۰	۳.۰۲	-۴۲.۹۴	۳۵.۴۴	۴۵.۹۶
گدنا	-۴.۰۴	۱۹.۸۱	-۲۶.۱۰	۲۲.۰۵	۴۵.۹۰
های وب	۱۹.۳۳	۱۷.۶۶	-۲۸.۰۳	۴۷.۳۶	۴۵.۷۰
خفناور	۳.۱۰	۰.۰۰	-۴۵.۳۴	۴۸.۴۴	۴۵.۳۴
خوساز	۸۸.۶۳	۴۶.۳۸	۱.۴۸	۸۷.۱۵	۴۴.۹۰
پردیس	-۷.۴۸	۶.۲۴	-۳۷.۶۳	۳۰.۱۵	۴۳.۸۷

فصلنامه مهندسی مالی و مدیریت اوراق بهادار / شماره چهل و هشتم / پاییز ۱۴۰۰

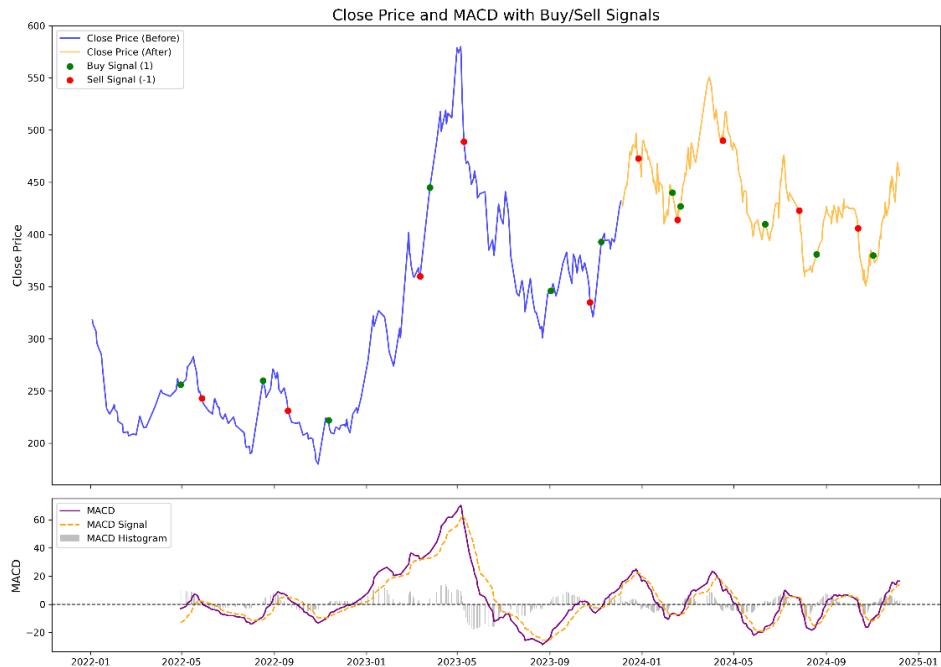


نمودار ۱- مقایسه بازدهی سه استراتژی در نمادهای پرتراکنش بورس تهران



نمودار ۲- نمادهایی با بیشترین اختلاف بازدهی استراتژی DBSCAN و خرید و نگهداری

استراتژی معاملاتی بر اساس خوشه بندی داده های قیمت.../ شریفی، رستمی



شکل ۶ - نمودار قیمتی نماد ذوب و داده های نویز با DBSCAN



شکل ۷- نمودار قیمتی نماد ذوب و داده های نویز با GSDM

بحث و نتیجه گیری

در این مطالعه، ما یک استراتژی معاملاتی مبتنی بر الگوریتم خوشه‌بندی DBSCAN و ویژگی‌های مهندسی شده توسعه دادیم و عملکرد آن را با استراتژی‌های ساده‌تر مانند تحلیل انحراف معیار گاوسی و استراتژی "خرید و نگهداری" مقایسه کردیم. نتایج حاصل از پژوهش نشان داد که به طور کلی استراتژی پیشنهادی ما در اکثر شرایط بازار عملکرد بهتری داشته و توانسته است بازدهی بیشتری نسبت به سایر روش‌ها ارائه دهد. یکی از نقاط قوت اصلی استفاده از DBSCAN این است که این الگوریتم به خوبی قادر است خوشه‌هایی با اشکال پیچیده و داده‌های پرت را شناسایی کند. این ویژگی باعث شده که استراتژی پیشنهادی بتواند از الگوهای پنهان در داده‌ها بهره‌برداری کرده و تصمیم‌گیری بهینه‌تری انجام دهد. در مقایسه، روش‌های مبتنی بر شاخص‌های آماری به دلیل فرضیات ساده‌سازی شده، مانند توزیع نرمال داده‌ها، در شناسایی الگوهای پیچیده بازار محدودیت داشته‌اند.

از سوی دیگر، استراتژی "خرید و نگهداری" علی‌رغم سادگی، به دلیل عدم انعطاف‌پذیری در مواجهه با تغییرات بازار و ریسک بالا به خصوص در دوره‌های نزولی و رنج بازار، بازدهی کمتری نسبت به استراتژی معرفی شده داشته است و این مقایسه نشان می‌دهد که بهره‌گیری از الگوریتم‌های پیشرفته می‌تواند در تحلیل بازارهای مالی به خصوص در دوره‌های رنج و نزولی بازار در بهبود تصمیم‌گیری در معاملات مؤثر باشد. لازم به ذکر است خروجی این استراتژی در بازارهای دوطرفه و با محدودیتهای کمتر نسبت به بازار ایران می‌تواند نتایج متفاوت و بهتری داشته باشد.

با توجه به نتایج به‌دست‌آمده، نتایج نشان داد که استراتژی مبتنی بر DBSCAN در مواجهه با شرایط مختلف بازار (روند صعودی، نزولی و خنثی) عملکرد پایدارتری دارد. این موضوع نشان‌دهنده توانایی این الگوریتم در شناسایی الگوهای نهفته در داده‌های مالی است.

همچنین با توجه به اینکه الگوریتم DBSCAN از داده‌های پرت برای شناسایی الگوهای نامعمول استفاده می‌کند، این موضوع می‌تواند به کاهش ریسک در استراتژی معاملاتی کمک کند. در مقابل، روش‌های ساده‌تر به دلیل نادیده گرفتن داده‌های پرت ممکن است فرصت‌های سودده را از دست بدهند. و در نهایت استفاده از الگوریتم‌های پیشرفته مانند DBSCAN نیازمند منابع محاسباتی بیشتری است که برای کاربرد در محیط‌های با داده‌های بزرگ، نیاز به بهینه‌سازی و استفاده از زیرساخت‌های مناسب خواهد بود.

پیشنهادهای برای تحقیقات آینده

تطبیق با الگوریتم‌های دیگر: بررسی عملکرد استراتژی پیشنهادی در ترکیب با الگوریتم‌های خوشه‌بندی دیگر مانند K-Means یا Gaussian Mixture Model می‌تواند به بهبود بیشتر نتایج کمک کند.

استراتژی معاملاتی بر اساس خوشه بندی داده های قیمت.../ شریفی، رستمی

گسترش دامنه داده‌ها: استفاده از داده‌های تاریخی بیشتر و اعمال این استراتژی در بازارهای مالی مختلف می‌تواند به ارزیابی بهتر کارایی روش کمک کند.

افزودن جنبه‌های زمانی: ترکیب ویژگی‌های زمانی، مثل زنجیره مارکوف یا مدل‌های ARIMA با الگوریتم DBSCAN ممکن است دقت پیش‌بینی را افزایش دهد.

این تحقیق نشان داد که استفاده از الگوریتم‌های یادگیری ماشین مانند DBSCAN در استراتژی‌های معاملاتی می‌تواند ابزار مؤثری برای تحلیل و پیش‌بینی بازارهای مالی باشد. در مقایسه با روش‌های سنتی و استراتژی‌های ساده مانند "خرید و نگهداری"، رویکرد پیشنهادی توانسته است بازدهی بهتری ارائه دهد. با این حال، برای بهره‌برداری کامل از این روش، انجام تحقیقات بیشتر در راستای بهینه‌سازی و گسترش آن پیشنهاد می‌شود.

منابع

- ۱) Abbondante, P. (۲۰۱۰). *Trading volume and stock indices: A test of technical analysis*. researchgate.net.
https://www.researchgate.net/profile/Paul-Abbondante-2/publication/49619617_Trading_Volume_and_Stock_Indices_A_Test_of_Technical_Analysis/links/5f170f7e45851515ef3bfe47/Trading-Volume-and-Stock-Indices-A-Test-of-Technical-Analysis.pdf?sg%0B%0D=started_experiment_milestone&origin=journalDetail
- ۲) Ahadzie, R., & Jeyasreedharan, N. (۲۰۲۰). Trading volume and realized higher-order moments in the Australian stock market. *Journal of Behavioral and Experimental*
<https://www.sciencedirect.com/science/article/pii/S2214635020303403>
- ۳) Ahmed, M., & Mahmood, A. (۲۰۱۳). A novel approach for outlier detection and clustering improvement. *2013 IEEE 8th Conference on*
<https://ieeexplore.ieee.org/abstract/document/6566435/>
- ۴) Attari, M., Rafiq, S., & Awan, H. (۲۰۱۲). *The dynamic relationship between stock volatility and trading volume*. academia.edu.
https://www.academia.edu/download/97406642/28_201085-1097.pdf
- ۵) Chatzikonstanti, V. (۲۰۱۷). Breaks and outliers when modelling the volatility of the US stock market. *Applied Economics*.
<https://doi.org/10.1080/00338446.2017.1293780>

- ۶) Chutka, J., & Rebetak, F. (۲۰۲۱). Analysis of trading volume and its use in prediction future price movements in the process of maximizing trading earnings. *SHS Web of Conferences*. https://www.shs-conferences.org/articles/shsconf/abs/۲۰۲۱/۰۳/shsconf_glob۲۰۲۰۱۰/shsconf_glob۲۰۲۰۱۰.html
- ۷) El-Ansary, O., & Atuea, M. (۲۰۱۲). *The effect of stock trading volume on return in the Egyptian Stock Market*. academia.edu. https://www.academia.edu/download/۴۱۷۲۵۲۲۹/The_Effect_of_Stock_Trading_Volume_on_Re۲۰۱۶۰۱۲۹-۱۵۲۲۵-۱nqsjow.pdf
- ۸) González-Sánchez, M. (۲۰۲۱). Is there a relationship between the time scaling property of asset returns and the outliers? Evidence from international financial markets. *Finance Research Letters*. <https://www.sciencedirect.com/science/article/pii/S۱۵۴۴۶۱۲۳۱۹۳۰۶۷۷۴>
- ۹) Nipane, S., Kale, P., Kapsekar, P., Agrawal, P., & Dudbhure, O. (۲۰۲۱). *Price trend prediction of stock market using outlier data mining algorithm*.
- ۱۰) Schluter, C., & Trede, M. (۲۰۰۸). Identifying multiple outliers in heavy-tailed distributions with an application to market crashes. *Journal of Empirical Finance*. <https://www.sciencedirect.com/science/article/pii/S۰۹۲۷۵۳۹۸۰۷۰۰۱۰۴۱>
- ۱۱) Shehadeh, A., Alwadi, S., & ... (۲۰۲۲). Detecting and Analysing Possible Outliers in Global Stock Market Returns. *Cogent Economics & ...*. <https://doi.org/۱۰.۱۰۸۰/۲۳۳۲۲۰۳۹,۲۰۲۲,۲۰۶۶۷۶۲>
- ۱۲) Suominen, M. (۲۰۰۰). Trading volume and information revelation in stock market. *Journal of Financial and Quantitative Analysis*. <https://www.cambridge.org/core/journals/journal-of-financial-and-quantitative-analysis/article/trading-volume-and-information-revelation-in-stock-market/۴EE۷۱۲۰A۷۵۴۱D۱۰B۵۸۰FC۵۵۵DEEB۱B۷۷>
- ۱۳) Tran, T., Drab, K., & Daszykowski, M. (۲۰۱۳). Revised DBSCAN algorithm to cluster data with dense adjacent clusters. *Chemometrics and Intelligent ...*. <https://www.sciencedirect.com/science/article/pii/S۰۱۶۹۷۴۳۹۱۲۰۰۲۲۴۹>
- ۱۴) Zhao, L., & Wang, L. (۲۰۱۵). Price trend prediction of stock market using outlier data mining algorithm. *2015 IEEE Fifth International Conference on ...*. <https://ieeexplore.ieee.org/abstract/document/۷۳۱۰۷۲۲/>

-
- ^۱ noise
 - ^۲ Nipane
 - ^۳ Skewness
 - ^۴ Kurtosis
 - ^۵ outlier data
 - ^۶ Ahadzie & Jeyasreedharan
 - ^۷ Shehadeh
 - ^۸ González-Sánchez
 - ^۹ Schluter & Trede
 - ^{۱۰} Machine Learning - ML
 - ^{۱۱} Density-Based Spatial Clustering of Applications with Noise - DBSCAN
 - ^{۱۲} CUSUM-Cumulative Sum
 - ^{۱۳} Chatzikonstanti
 - ^{۱۴} tick-by-tick
 - ^{۱۵} Heavy-Tailed Distribution
 - ^{۱۶} El-Ansary & Atuea
 - ^{۱۷} K-Means
 - ^{۱۸} Ahmed & Mahmood
 - ^{۱۹} Attari
 - ^{۲۰} Chutka & Rebetak
 - ^{۲۱} Abbondante
 - ^{۲۲} Suominen
 - ^{۲۳} Robust Clustering using Links - Rock
 - ^{۲۴} Tran
 - ^{۲۵} Eps
 - ^{۲۶} MinPts
 - ^{۲۷} Feature Engineering
 - ^{۲۸} Normalization
 - ^{۲۹} Random Forest
 - ^{۳۰} Correlation Analysis.
 - ^{۳۱} Random Forest
 - ^{۳۲} Feature Importance
 - ^{۳۳} Principal Component Analysis
 - ^{۳۴} Grid Search
 - ^{۳۵} Unseen Data
 - ^{۳۶} Moving Average Convergence Divergence
 - ^{۳۷} Gaussian Standard Deviation Method - GSDM
 - ^{۳۸} Buy and Hold

Developing a Trading Strategy Based on Price Data Clustering Using the DBSCAN Machine Learning Algorithm and Identifying Noise Data

Nastaran Sharifirad^۱

Mohammadreza Reza Rostami^۲

Receipt: Acceptance

Abstract

In this study, a trading strategy based on clustering price data using the DBSCAN machine learning algorithm is proposed and its performance is evaluated in the Tehran Stock Exchange. For this purpose, historical data of trading symbols in the stock markets were analyzed, and key features were extracted from price and volume data. The performance of this strategy was assessed using strategy return indicators and compared with a classical statistical method (based on normal distribution and standard deviation analysis) as well as the traditional "buy-and-hold" strategy.

The results demonstrated that the DBSCAN-based strategy, with its ability to identify complex clusters and outliers, outperformed the statistical method and provided higher returns. Moreover, this strategy achieved not only higher returns compared to the "buy-and-hold" strategy but also effectively reduced trading risks in range-bound and bearish markets. The findings of this research suggest that utilizing advanced clustering algorithms can serve as an effective tool for financial market analysis and trading decision-making in the Tehran Stock Exchange.

Keywords

Trading Strategy, Machine Learning, Clustering, Outlier Data

^۱.Master's Student, Department of Financial Management, Tehran Unit, Alzahra University, Iran. N.sharifirad@student.alzahra.ac.ir

^۲.Associate Professor, Department of Financial Management, Tehran Unit, Alzahra University, Iran. M.rostami@alzahra.ac.ir