فصلنامه راهبرد مدیریت مالی
دانشگاه الزهرا
سال دهم، شماره سی و هشتم، پاییز ۱۴۰۱
صفحات ۲۶-۱

مقاله پژوهشی

پیش بینی شاخص کل بورس اوراق بهادار تهران با استفاده از رگرسیون بردار پشتیبان بر مبنای تکنیک کاهش ابعاد [۱]

سمیه محبی [۲]، محمداسماعیل فدائی نژاد [۳]، محمد اصولیان [۴]

تاریخ دریافت: ۱۴۰۰/۰۱/۱۰ تاریخ پذیرش: ۱۴۰۱/۰۶/۱۹

چکیده

بازارهای سهام نقش مهمی در سازماندهی سیستم‌های اقتصادی مدرن دارند. پژوهش‌های گسترده‌ای در زمینه پیش‌بینی آن‌ها با استفاده از تکنیک‌های هوشمند انجام شده است. با توجه به این‌که دقت عملکرد این تکنیک‌ها به میزان قابل توجهی تحت تأثیر ویژگی‌های ورودی آن است، یکی از پیشرفت‌های به‌کار رفته در استفاده از مدل‌های هوشمند، علاوه‌بر کاربرد مدل‌های ترکیبی، استفاده از کاهش ابعاد به‌عنوان یک پیش‌مرحله برای مدل پیش‌بینی می‌باشد. در این پژوهش برای پیش‌بینی روزانه شاخص کل بورس اوراق بهادار تهران همزمان از دو روش کاهش ابعاد (انتخاب و استخراج) به منظور انتخاب ویژگی‌های مناسب به عنوان ورودی‌های مدل استفاده می‌شود. به‌طوری‌که برای انتخاب ویژگی‌ها از الگوریتم mRMR-MID و برای استخراج ویژگی‌ها از الگوریتم PCA استفاده می‌شود. سپس از رگرسیون بردار پشتیبان به‌عنوان مدل پیش‌بینی استفاده می‌شود. با توجه به نتایج بدست آمده از تحلیل استفاده از تکنیک‌های کاهش ابعاد در مدل پیش‌بینی، در نهایت الگوریتمی برای انتخاب ویژگی‌های مناسب بر شاخص، تحت عنوان ISF_MID پیشنهاد می‌شود. نتایج نشان می‌دهد که با روش پیشنهادی، می‌توان با ۷ ویژگی انتخابی به‌دقت بالایی در پیش‌بینی روزانه شاخص کل بورس اوراق بهادار تهران با درصد خطا ۳/۴۶ دست‌یافت. لازم به ذکر است مدل‌های مورد بررسی در مرحله پیاده‌سازی با روش اعتبارسنجی متقابل k-fold مورد ارزیابی قرار گرفتند. همچنین از معیارهای MAE، MSE و RMSE برای ارزیابی عملکرد مدل‌های مذکور استفاده می‌شود.

واژگان کلیدی: پیش‌بینی شاخص بورس، رگرسیون بردار پشتیبان، تکنیک کاهش ابعاد، انتخاب ویژگی، تجزیه و تحلیل مولفه‌های اصلی.
طبقه‌بندی موضوعی: *C38, C53, C61, G10, G17*

---

۱. کد DOI مقاله: 10.22051/JFM.2022.35543.2528
۲. دانشجوی دکتری، رشته مدیریت مالی، دانشکده مدیریت و حسابداری، دانشگاه شهید بهشتی تهران، ایران. E-mail: somayeh.mohebi@gmail.com.
۳. دانشیار، گروه مدیریت مالی و بیمه، دانشکده مدیریت و حسابداری، دانشگاه شهید بهشتی، تهران، ایران. نویسنده مسئول. E-mail: m-fadaei@sbu.ac.ir .
۴. استادیار، گروه مدیریت مالی و بیمه، دانشکده مدیریت و حسابداری، دانشگاه شهید بهشتی، تهران، ایران. E-mail: m_osoolian@sbu.ac.ir.

مقدمه

بازارهای سهام به‌دلیل داشتن پتانسیل‌های بالای مالی، توجه بسیاری از سرمایه‌گذاران را به خود جلب نموده است(نواسالمی[1]، ۲۰۲۰). از این‌رو پیش‌بینی این بازارها در میان تحلیل‌گران خبره و سرمایه‌گذاران از اهمیت بالایی برخوردار بوده و بسیار چالش برانگیز است (کومار، سارانجی و ورما[2]، ۲۰۲۱). این مسئله عمدتا به این دلیل است که بازار سهام یک سیستم بی‌نظم، پرآشوب، ناپارامتریک، پویا، غیرخطی و نامانا می‌باشد(رافیوزامن[3]، ۲۰۱۴).

برای پیش‌بینی و تحلیل بازارهای سهام از مدل‌های آماری و یادگیری ماشین استفاده شده است. مدل‌های آماری مانند میانگین متحرک خود رگرسیون انباشته[4](ARIMA) واریانس شرطی خودرگرسیونی تعمیم یافته[5](GARCH)، و نوسانات تصادفی[6] (SV)، هنوز هم برای پیش‌بینی سری‌های زمانی در اقتصاد استفاده می‌شوند، دلیل اصلی آن این است که خصوصیات ریاضی آن‌ها به‌خوبی درک شده است (الحق و همکاران[7]، ۲۰۲۱). با این حال، روش‌های آماری بازارهای سهام را خطی، ثابت و نرمال فرض می‌کنند که عملکرد و کاربرد عملی آن‌ها را در این حوزه محدود می‌کند. مدل‌های یادگیری ماشین مانند ماشین بردار پشتیبان[8](SVM)، شبکه‌های‌عصبی مصنوعی[9] (ANN)، سیستم‌های فازی[10](FIS) و الگوریتم ژنتیک[11](GA)، در سال‌های اخیر به‌منظور توضیح رفتار قیمت‌ها و پیش‌بینی آن‌ها در بازار سرمایه به دلیل تحلیل‌های چند متغیره بدون هیچ فرض از پیش تعیین شده‌ای با ویژگی غیرخطی و سهولت تعمیم، محبوبیت فراوانی یافته‌اند(ژونگ و انکه[12]، ۲۰۱۷). از میان این روش‌ها، SVM که یکی از روش‌های یادگیری با نظارت است به‌طور گسترده‌ای برای پیش‌بینی قیمت سهام مورد استفاده قرار گرفته است. استفاده از SVM برای پیش‌بینی تغییر قیمت روزانه در شاخص قیمت سهام ترکیبی کره[13] (KOSPI) که توسط کیم در سال ۲۰۰۳ انجام شد، هنوز هم یکی از پراستنادترین کارها در این زمینه است(بوستوس و پومارس-کویمبایا[14]، ۲۰۲۰). اخیرا هنریکه، سوبریرو و کیمورا[15] (۲۰۱۹)، ۵۴۷ مقاله را بررسی کردند و دریافتند که در ۳۷٪ مطالعات بررسی شده از SVM برای پیش‌بینی قیمت سهام استفاده نموده‌اند.

1. Nevasalmi
2. Kumar, Sarangi & Verma
3. Rafiuzzaman
4. Autoregressive Integrated Moving Average
5. Generalized Autoregressive Conditional Heteroskedasticity
6. Stochastic Volatility
7. Ul Haq et al.
8. Support Vector Machine
9. Artificial Neural Network
10. Fuzzy Inference System
11. Genetic Algorithm
12. Zhong & Enke
13. Korean composite stock price
14. Bustos & Pomares-Quimbaya
15. Henrique, Sobreiro & Kimura

هرچند طبق مطالعات انجام شده، ماشین بردار پشتیبان عملکرد جالب توجه و مناسبی در پیش‌بینی و طبقه‌بندی دارد، اما دقت عملکرد آن به‌میزان قابل توجهی تحت تأثیر نوع و تعداد ویژگی‌های ورودی آن است. بنابراین کاهش تعداد ویژگی‌هایی که باید در آموزش ماشین بردار پشتیبان به‌کار گرفته شود، تاثیر به‌سزایی در افزایش دقت نتایج و کاهش هزینه دارد(لی[1] ۲۰۰۹، ژانگ و همکاران[2] ۲۰۱۴). با توجه به این‌که بازارهای سهام تحت تاثیرعوامل مرتبط زیادی از جمله وضعیت اقتصادی، متغیرهای خاص صنعت، چشم‌انداز شرکت، نفوذ روانشناختی سرمایه‌گذاران و سیاست‌های دولت قرار دارند و بسیاری از آن‌ها به‌عنوان متغیرهای ورودی در طول توسعه یک سیستم پیش‌بینی بازار سهام ممکن است مورد استفاده قرار گیرند. بنابراین اگر از SVM انتظار می‌رودکه پیش‌بینی دقیق و کارآمدی را انجام دهد، ضروری است تا ویژگی‌هایی که حاوی مفیدترین اطلاعات باشند به‌عنوان ورودی‌های SVM انتخاب شوند. این نوع گزینش وظیفه اصلی تکنیک کاهش ابعاد[3] می‌باشد. کاهش ابعاد می‌تواند با دو شیوه متفاوت تحت عنوان انتخاب و استخراج ویژگی، اجرا شود(ژونگ و انکه،۲۰۱۷)، این دو روش اهمیت قابل توجهی در یادگیری ماشین دارند، زیرا منجر به کاهش پیچیدگی‌های محاسباتی شده و به بهبود توانایی تعمیم الگوریتم طبقه‌بندی کمک می‌کند. علت این امر کاهش اندازه نمونه‌های آموزشی می‌باشد که خطر "بیش برازش[4]" را کاهش می‌دهد(کاوالکانت و همکاران[5]،۲۰۱۶).

با وجود این‌که شناسایی و انتخاب ویژگی‌های ورودی با دقت پیش‌بینی بالا، مدت‌ها است که یک موضوع پژوهش در داده‌کاوی می‌باشد (الحق وهمکاران،۲۰۲۱)، و محققین در آمار، علوم کامپیوتر و ریاضیات کاربردی در این زمینه سال‌های زیادی کار کردند و انواع مختلفی از تکنیک‌های کاهش ابعاد خطی و غیرخطی را توسعه دادند، با این حال مطالعات موجود اتکا به استفاده از یک نوع تکنیک کاهش ابعاد (انتخاب یا استخراج ویژگی) را نشان می‌دهند. این امر ممکن است برخی از مفروضات مهم در مورد عملکرد رگرسیون اصلی متصل به متغیرهای ورودی و خروجی را نادیده بگیرد. بنابراین در این مطالعه، همزمان از دو شیوه متفاوت کاهش ابعاد استفاده می‌شود. به‌طوری‌که در روش انتخاب ویژگی از معیار کمینه‌ی افزونگی بیشینه‌ی وابستگی[6] mRMR استفاده می‌شود، که برخلاف پژوهش‌های صورت گرفته در این حوزه که از مقدار همبستگی ویژگی‌های ورودی نسبت به خروجی مدل برای انتخاب ویژگی‌های مؤثر استفاده می‌شود، این روش آماری؛ ویژگی‌های مؤثر را با توجه به بیشینه‌سازی معیار وابستگی آماری مجموعه ویژگی‌ها با ویژگی هدف، و کمینه کردن اطلاعات متقابل در بین مجموعه ویژگی‌های انتخابی، گزینش می‌کند. برای استخراج ویژگی‌ها از تجزیه و تحلیل مولفه‌های اصلی[7](PCA) که محبوب‌ترین تکنیک خطی برای کاهش ابعاد می‌باشد، استفاده می‌شود. سپس با ارزیابی عملکرد دو نوع تکنیک کاهش ابعاد، الگوریتمی به‌منظور

1. Lee
2. Zhang et al.
3. Dimensionality Reduction
4. Overfit
5. Cavalcante et al.
6. Minimum Redundancy Maximum Relevance
7. Principal Component Analysis

انتخاب ویژگی‌های مناسب با هدف فیلتر کردن متغیرهای ورودی بی‌ربط و اضافی برای کاهش پیچیدگی و بهبود دقت مدل پیش‌بینی پیشنهاد می‌گردد. در نهایت با مقایسه عملکرد روش‌های مختلف داده‌کاوی، روش مناسب برای مدل پیش‌بینی انتخاب می‌گردد و بر اساس این روش ویژگی‌های مهم به‌عنوان ورودی-های مدل پیش‌بینی روزانه شاخص کل بورس اوراق بهادار تهران شناسایی می‌گردند. بنابراین برای نخستین-بار این پژوهش یک فرآیند داده‌کاوی جامع را برای پیش‌بینی شاخص روزانه بورس بر اساس ۶۹ ویژگی اقتصادی و مالی انجام می‌دهد. سپس یک مدل دو مرحله‌ای با بیشترین دقت برای پیش‌بینی روزانه شاخص کل بورس اوراق بهادار تهران ارائه می‌دهد.

مبانی نظری و مروری بر پیشینه پژوهش

مبانی نظری

پیشرفت سریع و گسترده علم و فناوری و تلاش محققان برای اصلاح روش‌ها و مدل‌های پیش‌بینی، اخیرا موجب افزایش شایان توجه گرایش به استفاده از تکنیک‌های پیشرفته هوشمند، به‌منظور توضیح رفتار قیمت‌ها و پیش‌بینی آن‌ها در بازار سرمایه شده است. این مدل‌ها به‌طور عمده از توانایی احصای فرایندهای غیرخطی، نامانا و نویزی برخوردارند (باجلان، فلاح‌پور و دانا،۱۳۹۶). یکی از مهم‌ترین روش‌های یادگیری هوشمند که در سال‌های اخیر در بازارهای مالی بسیار مورد استفاده قرار گرفته و به نتایج مطلوبی نیز دست یافته است، ماشین بردار پشتیبان(SVM) می‌باشد.

SVM ها متداول‌ترین الگوریتم در میان الگوریتم‌های جداسازی خطی به دو دلیل می‌باشند. ۱) داده‌ها در این مدل‌ها می‌توانند بدون داشتن فرضیات قوی به‌کار گرفته شوند. ۲) در حالی‌که بیشتر مدل‌های شبکه عصبی مرسوم ریسک تجربی(خطای طبقه بندی نادرست) را به‌حداقل می‌رسانند، SVM ریسک‌های ساختاری (خطای تعمیم) را نیز به‌حداقل می‌رساند(کاوالکانت و همکاران،۲۰۱۶). وی ۱ (۲۰۱۲)، شهرت و کارآمدی مدل SVM را به واسطه فرمول ویژه تابع هدف محدب با محدودیت ضرایب لاگرانژ دانست. به‌گفته یوان ۲ (۲۰۱۳)، راه‌حل‌های SVM ممکن است در سطح جهانی بهینه باشد، در حالی‌که شبکه‌های عصبی مرسوم معمولاً راه‌حل بهینه محلی تولید می‌کنند حتی ممکن است در برخی مسائل محلی نیز با شکست روبه برو شوند.

SVM در واقع یک طبقه‌بندی‌کننده دو وجهی است که سعی دارد در دو طبقه، ابرصفحه‌ای ایجاد نماید که فاصله هر طبقه تا ابرصفحه حداکثر باشد. داده‌های نقطه‌ای که به ابر صفحه نزدیکترند، برای اندازه‌گیری این فاصله به‌کار می‌روند. این داده‌های نقطه‌ای، بردارهای پشتیبان نام دارند(منصورفر، غیور و خالق پرست اطهری، ۱۳۹۴). در این روش، ساخت مدل شامل دو مرحله آموزش و آزمایش می‌باشد. در انتهای فاز آموزش، قابلیت تعمیم مدل آموزش داده شده با استفاده از داده‌های آزمایش ارزیابی می‌شود (وی،۲۰۱۲). از این رو، طراحی SVM کاملا به استخراج زیر مجموعه‌ای از داده‌های آموزش (بردارهای پشتیبان) که توانایی نمایش دادن ویژگی‌های پایدار و ثابت داده‌ها را دارند، وابسته است (فلاح پور، گل ارضی و فتوره‌چیان، ۱۳۹۲).

1.Wei
2.Yuan

با توجه به توانایی تخمین غیرخطی SVM، از آن هم در طبقه‌بندی [1] (SVC) و هم مسائل رگرسیون [2] (SVR) استفاده می‌شود (گوا-کیانگ [3]،۲۰۱۱). با استناد به نتایج پژوهش‌های بسیاری، می‌توان بیان کرد که SVR به‌عنوان گزینه جایگزین شبکه‌های عصبی مصنوعی و نیز روش‌های آماری، در کارهای تشخیص الگو و پیش‌بینی سری‌های زمانی مالی به طور گسترده‌ای مورد استفاده قرار گرفته است (کاوالکانت و همکاران،۲۰۱۶). اما هنوز هم محدودیت‌هایی در الگوهای یادگیری SVR عمدتا در حوزه مالی، که در آن‌ها داده‌ها بسیار نویزی، بی‌ثبات و دارای ابعاد بالا هستند، وجود دارد (کارا ، بویاجی اوغلو و بایکان [4]،۲۰۱۱). بنابراین آماده‌سازی داده‌ها اولین قدم به‌منظور استفاده موفقیت آمیز هر روش هوشمند از جمله SVR است. به‌طوری‌که پس از تعریف متغیرهای ورودی و خروجی برای مدل‌سازی سری‌های زمانی مالی و به دست آوردن داده‌های ورودی مورد استفاده در آموزش، استفاده از برخی روش‌های پیش‌پردازش روی این داده‌ها ممکن است بسیار مفید باشد، زیرا می‌تواند منجر به بهبود دقت مدل یادگیری هوشمند گردد. یکی از مکانیسم‌های پیش‌پردازش داده‌ها، کاهش ابعاد می‌باشد. تعداد متغیرهای لازم برای اندازه‌گیری هر مشاهده اندازه " ابعاد داده" نامیده می شود. زمانی که داده‌ها دارای ابعاد زیاد هستند، علی‌رغم فرصت‌هایی که به‌وجود می‌آورند، ممکن است منجر به اطلاعات نامربوط و / یا زاید شوند، که نه تنها هزینه محاسباتی را افزایش می‌دهد بلکه می‌تواند عملکرد مدل پیش‌بینی را تضعیف نماید. بنابراین کاهش ابعاد تحلیل‌گران را قادر می‌سازد تا ویژگی‌های مناسب از یک مجموعه داده انتخاب و سپس از آن‌ها برای آموزش مدل پیش‌بینی استفاده کنند، که این امر می‌تواند منجر به کاهش پیچیدگی درمدل یادگیری و در نتیجه باعث بهبود عملکرد تعمیم الگوریتم طبقه‌بندی و دقت پیش‌بینی شود(کاوالکانت و همکاران،۲۰۱۶).

کاهش ابعاد می‌تواند با دو شیوه متفاوت اجرا شود:۱) با انتخاب مرتبط‌ترین متغیرها از مجموعه داده‌های اصلی که انتخاب ویژگی [5] نامیده می‌شود، ۲) با تولید یک گروه کوچکتر از متغیرهای جدید، که هرکدام ترکیب مشخصی از متغیرهای ورودی قدیمی‌تر می‌باشند، که استخراج ویژگی [6] نامیده می‌شوند(ژونگ و انکه،۲۰۱۷). به‌گفته وب (۲۰۰۳)، یک تمایز مهم در مورد انتخاب ویژگی و استخراج ویژگی وجود دارد. مکانیسم‌های انتخاب ویژگی که به آن‌ها انتخاب زیر مجموعه ویژگی نیز گفته می‌شود، متغیرهایی را که برای مدل‌سازی داده‌های آموخته شده نامربوط هستند، شناسایی می‌کنند، و برخلاف روش‌های مبتنی بر استخراج ویژگی، این نوع روش‌ها معنای اصلی ویژگی‌ها را بعد از کاهش حفظ می‌کنند. درصورتی‌که، مکانیسم‌های استخراج ویژگی سعی می‌کنند مجموعه ویژگی‌های اصلی را به فضای ویژگی‌هایی با ابعاد پایین انتقال دهند بدون این‌که ماهیت مسئله تغییر کند (کاوالکانت و همکاران،۲۰۱۶).

1. Classification
2. Regression
3. Guo-Qiang
4. Kara, Boyacioglu & Baykan
5. Feature Selection
6. Feature Extraction

انتخاب ویژگی، یکی از تکنیک‌هایی است که در مبحث یادگیری ماشین و همچنین شناسائی الگوهای آماری مطرح است. انتخاب ویژگی با تعیین زیرمجموعه‌ای از ویژگی‌های موجود که بیشترین اهمیت را برای مساله طبقه‌بندی دارند، ابعاد مجموعه اولیه ویژگی‌ها را کاهش می‌دهد(لیو و ژنگ[1]،۲۰۰۶). برای انتخاب ویژگی، راه‌حل‌ها و الگوریتم‌های فراوانی ارائه شده است. تمامی روش‌ها تلاش می‌کنند مجموعه‌ای از ویژگی‌ها را انتخاب کنند که علاوه بر توصیف کارآمد داده‌های ورودی، در عین حفظ دقت نتایج پیش‌بینی، متغیرهای نویزی و نامربوط را کاهش دهند. روش‌های مختلف انتخاب ویژگی را می‌توان به سه دسته؛ جستجوی کامل، جستجوی تصادفی و جستجوی مکاشفه‌ای[2] تقسیم‌بندی نمود. در جستجوی کامل تمام زیرمجموعه‌های ممکن برای یافتن جواب بهینه بررسی می‌شوند، بنابراین این احتمال وجود دارد که بر روی مجموعه داده‌های بزرگ غیرعملی باشد، زیرا بسیار زمان‌بر می‌باشد. در جستجوی تصادفی محدوده کمتری از فضای کل حالات بررسی می‌شود. در این روش‌ها پیدا شدن جواب بهینه به اندازه منابع موجود و زمان اجرای الگوریتم بستگی دارد. در جستجوی مکاشفه‌ای از بحث همبستگی بین ویژگی‌های ورودی با یکدیگر و نیز با ویژگی خروجی، برای انتخاب ویژگی‌ها استفاده می‌کنند، که در علم آمار با دو اصطلاح توضیح داده می‌شود. اولی بحث ارتباط[3] می‌باشد، که وابستگی بین ویژگی انتخابی به‌عنوان ورودی با خروجی مدل را نشان می‌دهد، و دومی افزونگی[4] می‌باشد، که وابستگی بین ویژگی انتخابی به‌عنوان ورودی با ویژگی‌هایی که تا کنون به‌عنوان ورودی مدل انتخاب شده است، را نشان می‌دهد. در این راستا از یک الگو دقیق استفاده می‌شود که به‌حداکثر شدن وابستگی بین ویژگی ورودی با ویژگی خروجی، و حداقل شدن وابستگی بین ویژگی که قرار است انتخاب شود با ویژگی‌هایی که تاکنون انتخاب شده‌اند، منجر می‌گردد.

در روش استخراج ویژگی، فضای یک مسئله به فضای دیگر نگاشت می‌یابد. هدف از نگاشت، انتقال یک مسئله از فضایی که شامل تعداد زیادی ویژگی می‌باشد به فضایی که تعداد ویژگی‌های کمتری داشته باشد، است. در این نگاشت ماهیت مسئله به‌هیچ عنوان تغییر نمی‌کند و صرفا در یک فضایی با ابعاد کمتر همان داده‌ها را خواهیم داشت. تجزیه و تحلیل مولفه‌های اصلی (PCA) قدیمی‌ترین و شناخته‌ترین روش آماری برای استخراج ویژگی‌های مهم از مجموعه داده‌هایی با ابعاد گسترده[5] است. این متدولوژی به کار پیرسون[6](۱۹۰۱)، برمی‌گردد و مبتنی بر ایده تعریف یک سیستم یا فضای هماهنگ جدید است که داده‌های خام می‌توانند با استفاده از داده‌های بسیار کمتری بیان شوند، بدون آن‌که اطلاعات قابل توجهی از دست برود.

پیشینه پژوهش

روند مطالعات در پژوهش‌های داخلی و خارجی، برتری روش ماشین بردار پشتیبان را به روش‌های پیشین، از جمله روش شبکه‌های عصبی و نیز روش‌های سنتی نشان می‌دهد. هرچند طبق مطالعات انجام شده، ماشین بردار پشتیبان، عملکرد جالب توجه و مناسبی در پیش‌بینی و طبقه‌بندی دارد، اما دقت عملکرد آن به‌میزان قابل توجهی

1. Lui & Zheng
2 .Heuristic
3. Relevance
4. Redandency
5. High- dimensional
6. Pearson

تحت تأثیر نوع و تعداد ویژگی ورودی آن است. به‌طوری‌که کاهش تعداد ویژگی‌هایی که باید در آموزش ماشین بردار پشتیبان به‌کار گرفته شود، تاثیر به‌سزایی در افزایش دقت نتایج و کاهش هزینه دارد. بنابراین، به تدریج روش‌های کاهش ابعاد برای مقابله با این محدودیت و بهبود عملکرد مدل‌های پیش‌بینی معرفی شدند. در ادامه به ذکر چند نمونه از مطالعاتی پرداخته می‌شود که اخیراً در داخل و خارج از کشور انجام گرفته و بیشترین ارتباط را با موضوع پژوهش دارند.

در حوزه مطالعات پیش‌بینی روند، باجلان، فلاح‌پور و دانا(۱۳۹۶)؛ مدلی برپایه‌ی ماشین بردار پشتیبان وزن‌دهی شده همراه با روش انتخاب ویژگی هیبرید که مرکب از یک بخش فیلتر کننده و یک بخش پوشش دهنده به‌منظور انتخاب زیرمجموعه‌ای بهینه از ویژگی‌ها می‌باشد، ارائه نمودند. براساس نتایج، این پژوهش نشان داد مدل VW-SVM همراه با انتخاب ویژگی F-SSFS عملکرد بهتری در پیش‌بینی قیمت سهم، نسبت به روش‌های موجود دارد.

فلاح پور، گل ارضی و فتوره چیان (۱۳۹۲) در پژوهشی تحت عنوان «پی بینی روند حرکتی قیمت سهام با استفاده از ماشین بردار پشتیبان برپایه الگوریتم ژنتیک در بورس اوراق بهادار تهران» اقدام به پیش‌بینی روند حرکت و تغییرات قیمت سهام در بازار بورس اوراق بهادار تهران نمودند. آن‌ها متغیرهای ورودی ماشین بردار پشتیبان را توسط الگوریتم ژنتیک، بهینه‌سازی نمودند. نتایج پژوهش آن‌ها نشان داد که ماشین بردار پشتیبان برپایه الگوریتم ژنتیک، دقت بسیار بیشتری در پیش‌بینی نسبت به ماشین بردار پشتیبان ساده دارد.

راعی، نیک عهد قصیرائی و حبیبی (۱۳۹۵)، در پژوهشی به‌منظور افزایش دقت پیش‌بینی شاخص بورس اوراق بهادار تهران ترکیبی از روش‌های آماری و هوش مصنوعی را به‌کار برده‌اند. در این پژوهش ابتدا از روش PCA برای پالایش اولیه داده‌ها استفاده شده است. سپس با استفاده از رگرسیون بردار پشتیبان بهینه شده به‌وسیله الگوریتم حرکت تجمعی ذرات، به پیش‌بینی شاخص اقدام شده است. نتایج بدست آمده نشان داد که پیش‌پردازش روی داده‌ها، خطای پیش‌بینی مدل را به‌طور قابل ملاحظه‌ای کاهش داده است.

نگوین و لو[1] (۲۰۱۴)، با ترکیب SOM[2] که یک الگوریتم خوشه‌بندی است، و f-SVM[3]، به پیش‌بینی قیمت سهام پرداختند. با توجه به نتایج به‌دست آمده مدل دو مرحله ای یادشده نسبت به مدل‌های دیگر مانند مدل RBN، مدل ترکیبی SOM با SVM و مدل ANFIS عملکرد بهتری داشته است. لی پینگ نی، ژی وی نی و گائو[4] (۲۰۱۱) در پژوهشی با عنوان "پیش‌بینی روند سهام بر اساس انتخاب ویژگی فراکتال[5] و ماشین بردار پشتیبان؛ برای پیش-بینی یک روز آتی شاخص قیمتی سهام، از رویکرد ترکیبی SVM با روش انتخاب ویژگی فراکتال که برای حل مشکلات غیرخطی مناسب است و در انتخاب بهینه تعداد ویژگی‌ها کمک می‌کند، استفاده کردند. نتایج نشان از عملکرد مناسب‌تر این روش نسبت به روش‌های مشابه داشته است.

1. Nguyen & Le
2. Self-Organizing Map (SOM)
3. Fuzzy –Support Vector Machines
4. Li-Ping Ni, Zhi-Wei Ni & Ya-Zhuo Gao
5. Fractal Feature Selection

اوو و وانگ[1] (۲۰۰۹)، برای پیش‌بینی روند حرکتی قیمت شاخص بازار سهام هنگ‌کنگ، از ۱۰ روش داده‌کاوی استفاده کردند. این روش‌ها شامل تجزیه و تحلیل افتراقی خطی، تحلیل افتراقی درجه دوم، الگوریتم-Kنزدیکترین همسایه، الگوریتم دسته‌بندی بیز براساس تابع برآورد کرنل، مدل لاجیت، طبقه‌بندی درختی، شبکه عصبی، طبقه‌بندی بیزی با فرایند گاوسی، ماشین بردار پشتیبان و حداقل مربعات ماشین بردار پشتیبان می‌باشند. نتایج تجربی نشان داد که عملکرد پیش‌بینی مدل‌های ماشین بردار پشتیبان و حداقل مربعات ماشین بردار پشتیبان بهتر از سایر مدل‌ها می‌باشد.

هوانگ[2] (۲۰۱۲)، از GA برای بهینه‌سازی پارامترهای SVR و انتخاب ورودی‌های مدل استفاده کرد. ژانگ و همکاران (۲۰۱۴) از یک روش انتخاب ویژگی تحت عنوان CFS[3] همراه با ۷ مدل پایه‌ای پیش‌بینی و ۱۸ ویژگی ورودی، برای پیش‌بینی روند تغییرات شاخص بورس شانگهای استفاده کردند. نتایج نشان داد که روش انتخاب ویژگی پیشنهادی آن‌ها از لحاظ دقت، عملکرد بسیار بهتری نسبت به سه روش انتخاب ویژگی پرکاربرد (PCA, CART, LASSO) دارد. سینگ و سریواستاوا[4] (۲۰۱۷) از PCA برای سرعت بخشیدن به آموزش مدل بدون از دست دادن دقت در صحت پیش‌بینی استفاده کردند. در این مقاله، آن‌ها قیمت سهام را با استفاده ترکیب دو تکنیک قدرتمند PCA و شبکه عصبی عمیق پیش‌بینی کردند، سپس، نتایج خود را با شبکه عصبی تابع پایه شعاعی مقایسه کردند و نشان دادند دقت مدل پیشنهادی ۴.۸٪ بهبود یافته است.

ژانگ و انکه (۲۰۱۷) به‌منظور پیش‌بینی بازده شاخص بورس، سه تکنیک کاهش ابعاد شامل تحلیل مولفه اصلی (PCA)، تحلیل مؤلفه اصلی مقاوم فازی (FRPCA) و تحلیل مولفه اصلی مبتنی بر کرنل(KPCA) را بر ۶۰ داده مالی و اقتصادی به‌منظور ساده‌سازی و آرایش مجدد ساختار اصلی مورد استفاده قرار دادند. آن‌ها نتیجه گرفتند که استفاده از PCA نه تنها مقدار داده‌های مورد نیاز برای آموزش مدل‌ها را ساده می‌کند، بلکه باعث افزایش دقت کلی پیش‌بینی‌ها می‌شود.

پرسش های پژوهش

این پژوهش به‌دنبال پاسخگویی به پرسش‌های زیر است:

۱. کدام یک از تکنیک‌های کاهش ابعاد (انتخاب یا استخراج ویژگی) تاثیر بیشتری در دقت عملکرد مدل SVR به منظور پیش‌بینی شاخص بورس اوراق بهادار تهران دارند؟

۲. برای تخمین mRMR کدام یک از الگوریتم‌های انتخاب ویژگی با نام‌های MID و FCD، مناسب می‌باشند؟

۳. الویت ویژگی‌های انتخاب شده توسط MID و FCD، به چه صورت است؟

۴. آیا الگوریتم پیشنهادی در این پژوهش در مقایسه با روش mRMR و PCA، عملکرد بهتری در انتخاب ویژگی‌های مناسب برای پیش‌بینی شاخص بورس اوراق بهادار تهران دارد؟

1. Ou & Wang
2. Huang
3. Causal Feature Selection
4. Singh & Srivastava

۵. چه ویژگی‌هایی برای ورودی مدل پیش‌بینی شاخص روزانه بورس اوراق بهادار تهران مناسب هستند و منجر به بهبود عملکرد مدل پیش‌بینی می‌شوند؟

روش‌شناسی پژوهش

در این بخش از پژوهش به‌صورت اجمالی مجموعه‌داده، آماده‌سازی مجموعه‌ی داده، تکنیک‌های مختلف کاهش ابعاد، رگرسیون بردار پشتیبان به‌عنوان مدل پیش‌بینی شاخص بورس و معیارهای ارزیابی عملکرد مدل پیش‌بینی، معرفی می‌شوند.

ایجاد مجموعه داده

نظر به‌اینکه افزایش تعداد داده‌ها در شبکه‌های عصبی مصنوعی موجب کسب نتایج دقیق‌تر می‌شود، لذا داده‌های مورد استفاده در پژوهش که از بررسی مطالعات پیشین و همچنین مبانی تئوریک استخراج گردیده‌اند، بصورت روزانه از تاریخ ۱۳۹۲/۱۰/۲۸ تا ۱۳۹۷/۵/۳۰ به مدت ۱۱۰۸روز جمع‌آوری شدند. برای جمع‌آوری داده‌ها از پایگاه‌های اینترنتی شرکت خدمات فناوری بورس تهران، بانک مرکزی جمهوری اسلامی ایران، بانک داده‌های اقتصادی و مالی مربوط به وزارت امور اقتصادی و دارایی و اوپک بهره گرفته شده است.

جدول ۱. لیست ویژگی‌های مورد بررسی برای پیش‌بینی شاخص کل بورس اوراق بهادار تهران

نام ویژگی	توصیف مختصر ویژگی
Tehran Exchange Price Index(TEPIX)	بازده شاخص کل در روز جاری و سه روز گذشته
Trading Volume(TV)	تغییر نسبی حجم معاملات درسه روز گذشته
Index of 50 more active Companies(IC50)	بازده شاخص ۵۰ شرکت فعال‌بورس درسه روز گذشته
Price Index of 50 more active Companies(PIC50)	بازده شاخص قیمت ۵۰ شرکت فعال‌بورس درسه روز گذشته
30 Largest Companies Index (LCI30)	بازده شاخص ۳۰ شرکت بزرگ درسه روز گذشته
Industrial Index(InI)	بازده شاخص صنعت درسه روز گذشته
Financial Index(FI)	بازده شاخص مالی درسه روز گذشته
Exponential Moving Average(EMA)	میانگین متحرک نمایی قیمت پایانی ۱۰، ۲۰، ۵۰ و ۲۰۰روز گذشته
Moving Average(MA)	میانگین متحرک قیمت پایانی ۵، ۱۰ ، ۳۰ روز گذشته
Relative Difference in Percentage (RDP)	تفاوت نسبی در درصد بازده شاخص در ۵، ۱۰،۱۵ و ۲۰ روز گذشته
Moving Price level Percentage (MPP)	شاخص سطح مقاومت قیمت[1] برای ۳۰ و ۱۲۰ روز گذشته
Price /Earning per share (P/E)	نسبت قیمت به سود با ۴وقفه و تغییرات نسبی ۳ روز گذشته آن
OPEC Oil Prices(OP)	قیمت نفت اوپک با ۴ وقفه و تغییرات نسبی ۳ روز گذشته آن
USD/IRR (DP)	قیمت دلار با ۴ وقفه و تغییرات نسبی ۳ روز گذشته آن
Euro/IRR (EP)	قیمت یورو با ۴ وقفه و تغییرات نسبی ۳ روز گذشته آن
Gold Coin Price (GP)	قیمت سکه طلا با ۴ وقفه و تغییرات نسبی ۳ روز گذشته آن

۱. مفهومی در تحلیل تکنیکی است که گرایش سهام و شاخص‌ها را در پشتیبانی از کاهش قیمت یا مقاومت در برابر افزایش قیمت، ارزیابی می‌کند. MPP بالا بیان‌گر سطح بیشتر مقاومت می باشد.

آماده‌سازی مجموعه‌داده

برای یکسان کردن مقیاس پارامترها، مرحله نرمال‌سازی انجام شده است. برای نرمال‌سازی پارامترها که شامل تعیین مقدار هر پارامتر در بازه‌ی بین صفر و یک است، از رابطه ساده ۱ استفاده می‌شود:

$$\eta_{i,j} = \frac{v_{i,j} - \min(V_j)}{\max(V_j) - \min(V_j)} \qquad (۱)$$

در این رابطه، V_j مجموعه‌ی مقادیر پارامتر jام و $v_{i,j} \in V_j$ که i شماره نمونه در مجموعه‌داده است. $\eta_{i,j}$ مقدار نرمال شده نمونه iام از پارامتر j است.

اولویت‌بندی ویژگی‌های مؤثر بر شاخص

در یادگیری ماشین، تعداد زیاد ورودی‌های مدل، باعث کاهش تعمیم مدل و افزایش سربار محاسباتی آن می‌شود. کاهش تعمیم، به معنی کاهش انطباق مدل با داده‌هایی است که آموزش مدل با آن‌ها انجام نشده است. بنابراین، کاهش تعداد ویژگی‌های ورودی، می‌تواند باعث افزایش دقت مدل پیش‌بینی گردد. الگوریتم‌های متعددی شامل الگوریتم‌های هوشمند و الگوریتم‌های مبتنی بر الگوهای آماری برای انتخاب ویژگی پیشنهاد شده‌اند. اغلب الگوریتم‌های آماری علاوه بر اینکه سرعت بالاتری در اجرا دارند می‌توانند الویت هر ویژگی را بصورت عددی مشخص کنند درصورتی‌که الگوریتم‌های هوشمند اغلب یک مجموعه از ویژگی‌ها را به‌عنوان ویژگی‌های مناسب انتخاب می‌کنند و ترتیبی از الویت‌ها ارائه نمی‌دهند. استفاده از مقدار همبستگی ویژگی‌های ورودی نسبت به خروجی مدل، یکی از روش‌های آماری انتخاب ویژگی برای مدل پیش‌بینی است. اما بهتر است انتخاب ویژگی بر اساس وابستگی بین داده‌ها انجام شود. بنابراین، به منظور انتخاب ویژگی‌های مناسب برای مدل پیش‌بینی معیاری نیاز است علاوه بر توجه به همبستگی هر ویژگی با مقدار شاخص آتی، همبستگی بین ویژگی‌های ورودی با یکدیگر نیز بررسی شود. در این راستا، ویژگی‌ها با توجه به بیشینه‌سازی معیار وابستگی آماری مجموعه ویژگی‌ها با ویژگی هدف و کمینه کردن اطلاعات متقابل[1] (MI) در بین مجموعه ویژگی‌های انتخابی، گزینش می‌شوند. MI بین دو ویژگی x و y با رابطه ۲ محاسبه می‌شود.

$$I(x;y) = \iint p(x,y) \log \frac{p(x)}{p(x)p(y)} dxdy \qquad (۲)$$

در رابطه فوق، $p(x)$، $p(y)$ و $p(x,y)$ به ترتیب، توابع چگالی احتمال متغیرهای x ، y و وقوع هم‌زمان آن‌ها می‌باشند.

از رابطه ۳ برای محاسبه زیرمجموعه‌ی ویژگی‌ها با بیش‌ترین مقدار وابستگی با ویژگی هدف، استفاده می‌شود.

$$\max V_I(S,h), \quad V_I = \frac{1}{|S|}\sum_{i \in S} I(i,h) \qquad (۳)$$

S زیرمجموعه‌ای از ویژگی‌های اولیه و V_I مقدار وابستگی S را با ویژگی هدف h نشان می‌دهد.

انتخاب زیرمجموعه براساس بیش‌ترین وابستگی، می‌تواند شامل ویژگی‌هایی باشد که خود آن‌ها همبستگی بالایی باهم داشته باشند. بر این اساس، شرط حداقل افزونگی، برای یافتن زیرمجموعه‌ای از ویژگی‌ها دارای کمترین همبستگی با یکدیگر، ارائه گردیده است و از طریق رابطه ۴ محاسبه می‌شود.

1 . Mutual Information

$$\min W_I(S), \quad W_I = \frac{1}{|S|^2}\sum_{i,j\in S} I(i,j) \qquad (۴)$$

W_I میانگین MI بین ویژگی‌ها در زیرمجموعه S است. با ترکیب دو رابطه ۳ و ۴، مفهومی تحت عنوان کمینه افزونگی بیشینه وابستگی (mRMR) معرفی گردیده است (پرز-رودریگز و همکاران[1]،۲۰۰۴). اساس mRMR با رابطه ۵ تعریف می‌شود:

$$\max \varphi(V,W), \quad \varphi = V - W \qquad (۵)$$

که $\varphi(V,W)$ عملگری برای ترکیب روابط بیشینه‌ی وابستگی و کمینه‌ی افزونگی است. انتخاب فهرست بهترین ویژگی‌ها برای مجموعه داده‌ای با تعداد زیادی ویژگی، به دلیل نمایی بودن تعداد محاسبه‌های لازم برای یافتن حداکثر مقدار $\varphi(V,W)$، غیرعملی است. به همین دلیل روش‌هایی برای تخمین mRMR معرفی گردیده‌اند. از متداول‌ترین آن‌ها، می‌توان به روش تفاضل اطلاعات متقابل[2] (MID) اشاره کرد(دینگ و پنگ[3]،۲۰۰۵). MID با رابطه ۶ محاسبه می‌شود. باید توجه داشت که برای استفاده از MID نیاز به گسسته‌سازی داده‌ها است. بدین منظور با توجه به مقادیر بین کمینه و بیشینه‌ی شاخص بورس، تعداد بازه‌ها تعیین می‌شود.

$$max_{i\in\varphi(S)}[I(i,h) - \frac{1}{|S|}\sum_{j\in S} I(i,j)] \qquad (۶)$$

همچنین، رابطه‌ی ۷ تحت عنوان FCD[4] برای تخمین mRMR برای متغیرهای پیوسته ارائه شده است ماندال و موخوپادیای[5]،۲۰۱۳)

$$max_{i\in\varphi(S)}[F(i,h) - \frac{1}{|S|}\sum_{j\in S} c(i,j)] \qquad (۷)$$

در رابطه فوق، $c(i,j)$ مقدار همبستگی دو ویژگی i و j را نشان می‌دهد. تابع F با رابطه ۸ تعریف شده‌است.

$$F(g_i,h) = [\sum_k n_k(\bar{g}_k - \bar{g})/(K-1)]/\sigma^2 \qquad (۸)$$

مقدار تابع F برای متغیر g_i، با توجه به K دسته‌ای که بر اساس h تعیین شده، محاسبه می‌شود. $\bar{g}$ میانگین مقدار g_i در همه نمونه‌ها و $\bar{g}_k$ میانگین مقدار در دسته kام است. همچنین σ^2 که واریانس تلفیقی است با رابطه ۹ محاسبه می‌شود.

$$\sigma^2 = \left[\sum_k (n_k - 1)\sigma_k^2\right]/(n-K) \qquad (۹)$$

در رابطه فوق، σ_k^2 و n_k واریانس و اندازه دسته kام است.

1. Perez-Rodriguez et al.
2. Mutual Information Difference
3. Ding &Peng
4. F-Test Correlation Difference
5. Mandal &Mukhopadhyay

تجزیه و تحلیل مولفه‌های اصلی (PCA)

تحلیل مولفه‌های اصلی، محبوب‌ترین تکنیک خطی برای کاهش ابعاد است، و به‌عنوان یکی از اولین تکنیک‌های چند متغیری، به‌دنبال ساخت یک نماینده با ابعاد کمتر از داده‌ها در عین حفظ واریانس حداکثری و ساختار کوواریانس داده‌ها است.

اساسا مولفه‌های اصلی، ترکیب‌های خطی از همه عوامل با ضرایب برابر با عناصر بردارهای ویژه[1] می‌باشند. مقادیر مختلف مولفه‌های اصلی می‌توانند نسبت‌های مختلف ساختار واریانس-کوواریانس داده‌ها را توضیح دهند. مقادیر ویژه[2] می‌توانند برای رتبه‌بندی بردارهای ویژه بر این اساس که چه مقدار از تغییرات داده‌ها توسط هر مولفه اصلی گرفته می‌شود، مورد استفاده واقع شوند. به‌طور دقیق، چنانچه $\lambda^* = \{\lambda_i^*\}_{i=1}^{M}$ نشان‌دهنده مقادیر ویژه ماتریس همبستگی corr (X) باشد به‌طوری که $\lambda_1^* \geq \lambda_2^* \geq \cdots \geq \lambda_M^*$. همچنین بردارهای $e_i^T = (e_{i1}, e_{i2}, \dots e_{iM})$ نشان‌دهنده بردارهای ویژه‌ای از corr (X) مطابق با مقادیر ویژه$\lambda_i^*, i = 1,2 \dots, M$ می‌باشند. عناصر این بردارهای ویژه، ضرایب مولفه‌های اصلی هستند. یعنی، مولفه‌های اصلی داده‌های استاندارد شده عبارت است از $Z = (Z_1, Z_2, \dots, Z_M)$ به طوری که؛

$$Z_W^T = (Z_{1W}, Z_{2W}, \dots, Z_{NW}), Z_{VW} = \frac{X_{VW} - \mu_W}{\sigma_W}, V = 1,2, \dots, N \; and \; W = 1,2, \dots, M \quad (۱۰)$$

بنابراین، رابطه ۱۰ می‌تواند به این شکل ذیل نوشته شود:

$$Y_i = \sum_{j=1}^{M} e_{ij} Z_j, \quad i = 1,2, \dots, M \quad (۱۱)$$

و اثبات شده است که:

$$var(Y_i) = \sum_{k=1,l=1}^{M} e_{ik} \, corr(X_k, X_l) e_{il} = e_i^T \rho e_i = \lambda_i^* \quad (۱۲)$$

و

$$cov(Y_i, Y_j) = \sum_{k=1,l=1}^{M} e_{ik} \, corr(X_k, X_l) e_{jl} = e_i^T \rho e_j = 0 \quad (۱۳)$$

در نهایت، با استفاده از برهان تجزیه طیفی، می‌توان گفت؛

$$\rho = \sum_{i=1}^{M} \lambda_i^* e_i \, e_i^T \quad (۱۴)$$

و این واقعیت که هر دو $e_i^T e_i = \sum_{j=1}^{M} e_{ij}^2 = 1$ و بردارهای ویژه مختلف نسبت به یکدیگر عمود هستند به‌طوری‌که $e_i^T e_j = 0$، بنابراین واریانس i^{th}(بزرگترین)مولفه اصلی برابر باi^{th} بزرگترین مقدار ویژه است و مولفه‌های اصلی با یکدیگر ناهمبسته هستند.از آنجا که تغییرات کل Z به عنوان ردیابی ماتریس همبستگی ρ تعریف می شود، یعنی$trace(\rho) = \sum_{i=1}^{M} \lambda_i^*$، نسب تغییرات توضیح شده توسط i^{th} مولفه اصلی به شکل$\lambda_i^* / trace(\rho)$ تعریف می‌گردد که $i = 1,2, \dots m$. نسبت تغییرات توضیح شده توسط k

1 .Eigenvectors
2. Eigenvalues

اول مولفه اصلی به عنوان مجموع k اول مقادیر ویژه که توسط $trace(\rho)$ تفکیک شده، تعریف می شود، یعنی $\sum_{i=1}^{k}\lambda_i^* / \sum_{i=1}^{M}\lambda_i^*$. به لحاظ نظری، اگر نسبت تغییراتی که توسط k اول مولفه‌های اصلی توضیح داده می شود بزرگ باشد، با کاهش ابعاد فضای داده از M به k، اطلاعات زیادی از دست نمی‌رود.

برای تعیین این‌که چه تعداد و کدام مولفه‌های اصلی باید به‌عنوان ورودی‌ها برای طبقه‌بندی مورد استفاده قرار گیرند، لازم است توازن بین دقت پیش‌بینی مورد انتظار، هزینه (زمان و ....) و پیچیدگی سیستم در نظر گرفته شود. به‌این معنی که مولفه‌های اصلی انتخابی، باید داده‌ها را به‌بهترین شکل، در حالی که ساختار داده‌ها را تا حد امکان ساده کرده است، توضیح دهند.

رگرسیون بردار پشتیبان (SVR)

یکی از مدل‌های متداول برای توابع تخمین، رگرسیون بردار پشتیبان است، که برای حل مسائل رگرسیون غیرخطی مناسب است. در ادامه توضیحات مختصری در رابطه با نحوه‌ی عملکرد مدل ارائه می‌گردد. مجموعه داده M را در نظر بگیرید که شامل بردارهای ورودی x_i و خروجی متناظر y_i است. تعداد نمونه‌های این مجموعه داده برابر n است.

$$M = \begin{Bmatrix} (x_i, y_i) |\ i = 1,2, \dots, n \\ , x_i \in R^N , y_i \in R \end{Bmatrix} \quad (۱۵)$$

در رگرسیون، به دنبال تخمین تابع $f(x_i)$ هستیم که خروجی‌های آن حداقل فاصله را با مقادیر y_i داشته باشد. در رابطه ۱۶، δ خطای تصادفی با توزیع $N(0, \sigma^2)$ است.

$$y_i = f(x_i) + \delta \quad (۱۶)$$

برای حل یک مسئله رگرسیون غیرخطی با SVR، ابتدا با بهره‌گیری از رابطه‌ی ۱۷، ورودی‌ها به‌صورت غیرخطی به فضای ویژگی f با ابعاد زیاد که به‌صورت خطی با خروجی وابستگی دارند، نگاشت می‌شوند. در این رابطه، w بردار وزن، b مقدار بایاس و $\varphi(x)$ تابعی است که ورودی‌ها را از فضای R به فضای $R^{N\times h}$ تصویر می‌کند.

$$f(x_i) = w\varphi(x_i) + b \ \ |w \in R^{N\times h}, b \in R \quad (۱۷)$$

تابع خطی f به‌گونه‌ای است که به اندازه ε از مقادیر واقعی انحراف دارد. برای داده‌هایی که خارج از باند باشند، از تابع جریمه‌ای استفاده می‌شود که با رابطه‌ی ۱۸ ارائه گردیده است.

$$L_\varepsilon\big(y_i, f(x_i)\big) = \begin{cases} 0 & |y_i - f(x_i)| \le \varepsilon \\ |y_i - f(x_i)| - \varepsilon & otherwise \end{cases} \quad (۱۸)$$

برای محاسبه‌ی ریسک عملیاتی تابع f از رابطه‌ی ۱۹ استفاده می‌شود.

$$R_{emp}[f] = \sum_1^n L_\varepsilon\big(y_i, f(x_i)\big) \quad (۱۹)$$

برای کاهش ریسک عملیاتی مربوط به داده‌ها با توجه به بهینه‌سازی تابع رگرسیون از رابطه ۲۰ بهره گرفته می‌شود.

$$J = \frac{1}{2}\|w\|^2 + CR_{emp}[f] \quad (۲۰)$$

برای داده‌هایی که مقدار $|y - f(x_i)|$ آن‌ها بیشتر از ε باشد، مقدار ξ_i^+ و یا ξ_i^- که مقدار تخطی را نشان می‌دهد، با رابطه‌های ۲۱ و ۲۲ محاسبه می‌شود.

$$\xi_i^+ = y - f(x_i) - \varepsilon \quad (۲۱)$$

$$\xi_i^- = \varepsilon - y - f(x_i) \quad (۲۲)$$

همچنین، رابطه‌ی ۲۳ بین تابع جریمه و مقدارهای تخطی وجود دارد.

$$L_\varepsilon(y_i, f(x_i)) = \xi_i^+ + \xi_i^- \quad (۲۳)$$

برای تخمین تابع f از تابع هدف با رابطه‌ی ۲۴ استفاده می‌شود.

$$minimize\ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i^+ + \xi_i^-) \quad (۲۴)$$

S.t. $\forall\, i$

$-y_i + f(x_i) + \varepsilon + \xi_i^+ \geq 0$

$y_i - f(x_i) + \varepsilon + \xi_i^- \geq 0$

$\xi_i^+, \xi_i^- \geq 0$

سپس، از ضریب لاگرانژ برای ایجاد شکل دوگان رابطه ۲۴ استفاده شده و ساده‌سازی انجام می‌شود. در صورتی که α_i^+ و α_i^-، به ترتیب ضریب قیدهای اول و دوم رابطه۲۴ باشند، بعد از ساده‌سازی، رابطه ۲۵ حاصل می‌شود.

$$minimize\ \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i^+ - \alpha_i^-)(\alpha_j^+ - \alpha_j^-) < \varphi(x_i)\cdot\varphi(x_j) > -\sum_{i=1}^{n}(\alpha_i^+ - \alpha_i^-)y_i + \varepsilon\sum_{i=1}^{n}(\alpha_i^+ + \alpha_i^-) \quad (۲۵)$$

$$\text{S.t. } \begin{Bmatrix} \sum_{i=1}^{n}(\alpha_i^+ - \alpha_i^-) = 0 \\ \alpha.\alpha^* \in [0, C] \end{Bmatrix}$$

در مسائل غیرخطی، ضرب داخلی دو تابع $\varphi(x_i)$ و $\varphi(x_j)$ با تابع کرنل گوسی، در رابطه ۲۶ جایگزین می‌شود.

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \quad (۲۶)$$

درنهایت، تابع f با رابطه ۲۷ محاسبه می‌شود.

$$f(x) = \sum_{i=1}^{n}(\alpha_i^+ - \alpha_i^-)K(x_i, x) + b \quad (۲۷)$$

در این رابطه، محاسبه b با فرمول ۲۸ انجام می‌شود. SV بردار پشتیبان می‌باشد.

$$b = \frac{1}{n}\left\{\sum_{0<\alpha_i^-<C}\left[y_i - \sum_{x_j\in SV}(\alpha_j^+ - \alpha_j^-)K(x_i, x_j) - \varepsilon\right] + \sum_{0<\alpha_i^+<C}\left[y_i - \sum_{x_j\in SV}(\alpha_j^+ - \alpha_j^-)K(x_i, x_j) + \varepsilon\right]\right\} \quad (۲۸)$$

معیارهای ارزیابی مدل‌های پیش‌بینی

روش‌های متنوعی برای ارزیابی مدل در مرحله‌ی پیاده‌سازی وجود دارد. این مقاله، از روش اعتبارسنجی متقابل k-fold[1]، برای انتخاب مجموعه‌های آموزش و آزمایش بهره برده است. در این رویکرد k مرحله‌ای، مجموعه داده به ۱۰دسته متمایز تقسیم‌بندی می‌شوند. در هر مرحله، ۱-۱۰ دسته برای آموزش مدل و یک دسته برای آزمایش مدل انتخاب می‌شوند. البته، پیش از تقسیم‌بندی داده‌ها، ابتدا جایگشتی بر روی مجموعه‌داده انجام شده تا نمونه‌ها به‌هم‌ریخته شوند. این به‌هم‌ریختگی باعث تنوع نمونه‌ها در توزیع داده‌ها می‌شود.

در این پژوهش، برای بررسی دقت مدل‌های پیش‌بینی، از معیارهای میانگین قدر مطلق خطا[2] (MAE)، میانگین مربع خطا[3] (MSE)، و ریشه میانگین مربع خطا[4] (RMSE) استفاده شده است. معیار MAE با رابطه‌ی ۲۹ محاسبه می‌شود.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|t_i - y_i| \qquad (۲۹)$$

N تعداد نمونه‌ها، t_i مقدار هدف و y_i مقدار پیش‌بینی‌شده برای نمونه‌ی iام است.

همچنین، MSE با رابطه‌ی ۳۰ محاسبه می‌شود.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(t_i - y_i)^2 \qquad (۳۰)$$

درنهایت، RMSE با رابطه‌ی ۳۱ محاسبه می‌گردد.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(t_i - y_i)^2} \qquad (۳۱)$$

تجزیه و تحلیل داده‌ها

انتخاب ویژگی‌های مؤثر بر شاخص با استفاده از روابط تخمین mRMR

از مزیت‌های استفاده از روابط تخمین mRMR می‌توان به این نکته اشاره کرد که با آن‌که تخمین قابل‌قبولی از mRMR دارند، ولی پیچیدگی محاسباتی آن‌ها کم است. به بیانی دیگر، سربار پایین، سرعت و قابلیت اطمینان بالا، باعث استفاده از این روش‌های تخمین برای انتخاب ویژگی‌های مناسب گردیده است. روش‌های آماری مورد استفاده برای تخمین mRMR در این مقاله، MID و FCD می‌باشند. بنابراین ویژگی‌های موجود در مجموعه‌داده ورودی را یک‌بار با استفاده از الگوریتم MID و بار دیگر با استفاده از FCD، در بردار ζ بر اساس میزان وابستگی ویژگی‌ها با مقدار خروجی اولویت‌بندی می‌شوند. سپس در $|\zeta|$ گام، زیرمجموعه‌هایی شامل $1 \leq j \leq |\zeta|$ عنصر ابتدایی از ζ را در مدل پیشنهادی ارزیابی کرده و میزان

1. K-fold Cross Validation
2. Mean Absolute Error
3. Mean Square Error
4. Root Mean Square Error

خطای پیش‌بینی برای هر یک اندازه‌گیری می‌گردد. ترتیب اولویت ویژگی‌ها با بهره‌گیری از MID و FCD به ترتیب در جدول‌های ۲ و ۳ نمایش داده شده است.

جدول۲. اولویت‌بندی ویژگی‌ها با الگوریتم MID با توجه به شاخص بورس یک روز آینده

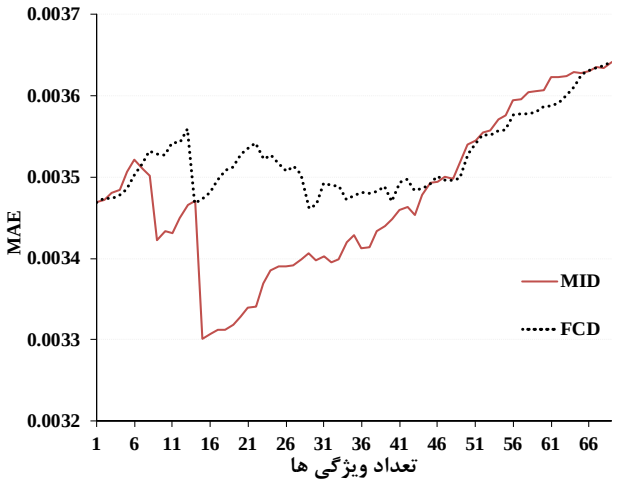
الویت	نام ویژگی	الویت	نام ویژگی	الویت	نام ویژگی	الویت	نام ویژگی	الویت	نام ویژگی
۱	RDP5	۱۵	$IC50_2$	۲۹	U_3	۴۳	InI_2	۵۷	GP_{d4}
۲	TV_1	۱۶	OP_1	۳۰	OP_{d4}	۴۴	MPP30	۵۸	P/E_{d2}
۳	P/E_{d3}	۱۷	$PIC50_1$	۳۱	P/E_2	۴۵	D_{d3}	۵۹	U_{d2}
۴	D_1	۱۸	U_1	۳۲	$PIC50_3$	۴۶	$TEPIX_3$	۶۰	D_{d1}
۵	$LCI30_1$	۱۹	FI_3	۳۳	U_2	۴۷	EMA_{20}	۶۱	OP_{d1}
۶	RDP10	۲۰	OP_3	۳۴	MPP120	۴۸	$TEPIX_2$	۶۲	GP_{d3}
۷	GP_2	۲۱	P/E_1	۳۵	InI_1	۴۹	OP_{d3}	۶۳	D_{d4}
۸	OP_2	۲۲	D_3	۳۶	InI_3	۵۰	P/E_{d4}	۶۴	EMA10
۹	$LCI30_3$	۲۳	RDP20	۳۷	$LCI30_2$	۵۱	GP_{d1}	۶۵	MA30
۱۰	GP_3	۲۴	FI_2	۳۸	GP_{d2}	۵۲	U_{d1}	۶۶	MA5
۱۱	FI_1	۲۵	P/E_3	۳۹	$PIC50_2$	۵۳	D_{d2}	۶۷	MA10
۱۲	GP_1	۲۶	TV_3	۴۰	$IC50_1$	۵۴	P/E_{d1}	۶۸	EMA50
۱۳	RDP15	۲۷	D_2	۴۱	U_{d4}	۵۵	U_{d3}	۶۹	EMA200
۱۴	TV_2	۲۸	$TEPIX_1$	۴۲	$IC50_3$	۵۶	OP_{d2}		

جدول۳. اولویت‌بندی ویژگی‌ها با الگوریتم FCD با توجه به شاخص بورس یک روز آینده

الویت	نام ویژگی	الویت	نام ویژگی	الویت	نام ویژگی	الویت	نام ویژگی	الویت	نام ویژگی
۱	RDP5	۱۵	GP_{d2}	۲۹	$TEPIX_3$	۴۳	MA30	۵۷	D_1
۲	RDP10	۱۶	GP_{d3}	۳۰	$LCI30_2$	۴۴	FI_3	۵۸	P/E_{d4}
۳	$IC50_1$	۱۷	D_{d3}	۳۱	$IC50_2$	۴۵	P/E_2	۵۹	U_1
۴	InI_1	۱۸	D_{d4}	۳۲	InI_2	۴۶	GP_3	۶۰	P/E_{d3}
۵	$PIC50_1$	۱۹	U_{d1}	۳۳	$TEPIX_2$	۴۷	GP_2	۶۱	$P/E_{d۲}$
۶	$LCI30_1$	۲۰	GP_{d4}	۳۴	EMA200	۴۸	FI_2	۶۲	GP_1
۷	RDP15	۲۱	U_{d2}	۳۵	U_2	۴۹	D_2	۶۳	OP_3
۸	RDP20	۲۲	U_{d3}	۳۶	P/E_3	۵۰	OP_1	۶۴	OP_2
۹	P/E_1	۲۳	$PIC50_3$	۳۷	MPP120	۵۱	OP_{d2}	۶۵	TV_2
۱۰	FI_1	۲۴	U_{d4}	۳۸	MA5	۵۲	OP_{d3}	۶۶	P/E_{d1}
۱۱	D_{d1}	۲۵	$PIC50_2$	۳۹	EMA10	۵۳	OP_{d4}	۶۷	TV_3
۱۲	MPP30	۲۶	$LCI30_3$	۴۰	MA10	۵۴	U_3	۶۸	TV_1
۱۳	D_{d2}	۲۷	$IC50_3$	۴۱	EMA20	۵۵	OP_{d1}	۶۹	$TEPIX_1$
۱۴	GP_{d1}	۲۸	InI_3	۴۲	EMA50	۵۶	D_3		

لازم به ذکر است که به منظور تقسیم داده‌ها در دو دسته آموزش و آزمایش از روش اعتبارسنجی متقابل k-fold استفاده گردید. در آزمایش انجام‌شده مقدار K برابر با ۱۰ در نظر گرفته شده است. علت

انتخاب مقدار ۱۰ با توجه به متداول بودن استفاده از این عدد توسط محققان می‌باشد. در این حالت، هر نمونه از داده‌ها ۹ بار برای آموزش و یک‌بار برای آزمایش مدل مورد استفاده قرار می‌گیرد. مقدار MAE در نمودار ۱ برای پیش‌بینی شاخص بورس یک روز آینده، برای هر زیرمجموعه از ویژگی‌های انتخاب‌شده نمایش داده شده است.



نمودار۱. استفاده از مدل SVR با ویژگی‌های اولویت‌بندی شده با روش‌های MID و FCD

با استفاده از این نمودار می‌توان مقایسه‌ای بین بهره‌گیری از MID و FCD در اولویت‌بندی ویژگی‌ها و اثر افزایش تعداد ویژگی‌های مدل پیش‌بینی داشت. یکی از نتایج قابل مشاهده در شکل این است که با افزایش تعداد ویژگی‌های اولویت‌بندی شده با FCD خطای مدل افزایش می‌یابد. در صورتی‌که با بهره‌گیری از اولویت‌بندی با MID، افزایش تعداد ویژگی‌ها ابتدا باعث افزایش دقت مدل می‌گردد ولی از ویژگی ۱۵ به بعد با کاهش دقت مدل روبه‌رو هستیم.

جدول ۴. مقایسه انتخاب ویژگی‌ها با استفاده از ۲ روش MID و FCD

۱۵ ویژگی اول	MID	FCD
درصد خطا	۳/۵۵	۳/۷۳
MAE	۰/۰۰۳۳	۰/۰۰۳۴
MSE	۰/۰۰۰۰۲۹	۰/۰۰۰۰۳۲
RMSE	۰/۰۰۵۳	۰/۰۰۵۶

نام ۱۵ویژگی ابتدایی که توسط MID الویت‌بندی شده و بالاترین دقت را در مدل SVR موجب می‌شوند، در جدول ۵ آمده است. از دیگر نکات قابل ملاحظه این است که اثر دو ویژگی ۹ و ۱۵ یعنی بازده

شاخص ۳۰ شرکت بزرگ در۳ روز گذشته و بازده شاخص ۵۰ شرکت فعال‌بورس در ۲ روز گذشته، تأثیر قابل توجهی در بهبود عملکرد مدل SVR داشته‌اند. نتایج در جدول ۶ ارائه شده است.

با مقایسه نتایج حاصل از MID و FCD، می‌توان گفت MID نتیجه بهتری در الویت‌بندی ویژگی‌ها ارائه نموده است. براین‌اساس، الگوریتم MID برای اولویت‌بندی و انتخاب ویژگی‌های مناسب برای مدل SVR به‌منظور پیش‌بینی شاخص بورس انتخاب گردید.

جدول ۵. پانزده ویژگی ابتدایی الویت‌بندی شده توسط MID

شماره ویژگی	۱	۲	۳	۴	۵
نام ویژگی	تفاوت نسبی درصد بازده شاخص در ۵ روز گذشته ($RDP5_{}$)	تغییر نسبی حجم معاملات در۱روز گذشته (TV_1)	نسبت قیمت به سود هرسهم در ۳ روز گذشته (P/E_{d3})	تغییر نسبی در قیمت دلار در۱ روز گذشته (D_1)	بازده شاخص ۳۰ شرکت بزرگ در ۱ روز گذشته ($LCI30_1$)
شماره ویژگی	۶	۷	۸	۹	۱۰
نام ویژگی	تفاوت نسبی درصد بازده شاخص در ۱۰ روز گذشته (RDP10)	تغییر نسبی در قیمت سکه در ۲روز گذشته (GP_2)	تغییر نسبی در قیمت نفت در ۲روز گذشته (OP_2)	بازده شاخص ۳۰ شرکت بزرگ در۳ روز گذشته ($LCI30_3$)	تغییر نسبی در قیمت سکه در ۳روز گذشته (GP_3)
شماره ویژگی	۱۱	۱۲	۱۳	۱۴	۱۵
نام ویژگی	بازده شاخص مالی در۱ روز گذشته (FI_1)	تغییر نسبی در قیمت سکه در۱ روز گذشته (GP_1)	تفاوت نسبی درصد بازده شاخص در ۱۵ روز گذشته (RDP15)	تغییر نسبی حجم معاملات در۲روز گذشته (TV_2)	بازده شاخص ۵۰ شرکت فعال‌بورس در ۲ روز گذشته ($IC50_2$)

جدول ۶. اثر دو ویژگی ۹ و ۱۵ بر عملکرد مدل SVR با ویژگی‌های اولویت‌بندی شده توسط الگوریتمMID

شماره ویژگی	نام ویژگی	درصد خطا	MAE	MSE	RMSE
شماره ۹	($LCI30_3$)	۳/۶۷	۰/۰۰۳۴۲۲	۰/۰۰۰۰۳۰	۰/۰۰۵۴۸۷
شماره ۱۵	($IC50_2$)	۳/۵۵	۰/۰۰۳۳۰۱	۰/۰۰۰۰۲۹	۰/۰۰۵۳۳۹

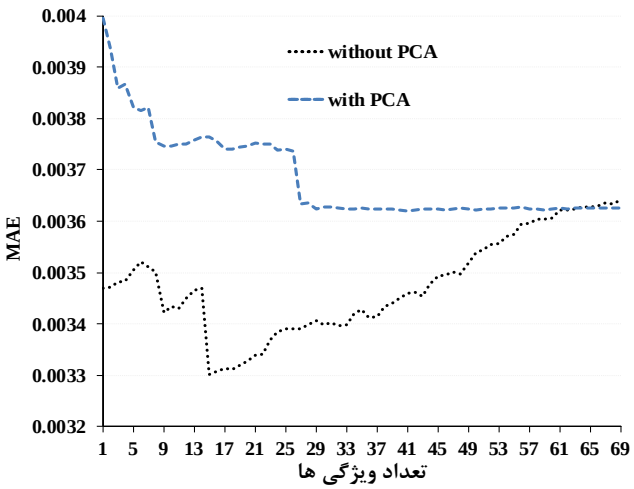
استخراج ویژگی‌های مؤثر بر شاخص با استفاده از تجزیه و تحلیل مولفه‌های اصلی (PCA)

در این مطالعه، به‌منظور استخراج ویژگی‌های جدید از PCA استفاده می‌شود. در این راستا، ابتدا همه ویژگی‌ها به الگوریتم PCA داده شده و با توجه به تعداد ابعاد خروجی که برای PCA در نظر گرفته می‌شود استخراج ویژگی‌ها انجام می‌شود. به عنوان مثال، اگر دو بعد به عنوان خروجی‌های PCA تعریف گردد، PCA همه‌ی ویژگی‌ها را دریافت کرده و با استفاده از الگوریتم مربوطه دو ویژگی استخراج می‌کند. برای این‌که بتوان راحت‌تر مقایسه استفاده و عدم استفاده از PCA را مشاهده کرد، نتیجه بهره‌گیری از هر دو تکنیک کاهش ابعاد، در نمودار۲ نمایش داده شده است. با توجه به نتایج مشاهده می‌شود که با افزایش

تعداد ابعاد PCA تا ۲۹ بعد، دقت مدل پیش‌بینی افزایش می‌یابد و بعد از آن تغییر تعداد ابعاد ویژگی‌ها بهبودی در دقت مدل ایجاد نمی‌کند. نتایج در جدول ۷ ارائه شده است. با مقایسه دو منحنی در نمودار ۲، دقت بالاتر استفاده از روش انتخاب ویژگی‌ها نسبت به استخراج ویژگی‌ها کاملا مشهود است. بنابراین می‌توان به این نتیجه رسید که بدون استفاده از PCA، عملکرد مدل SVR بهتر می‌باشد. بر این اساس، نیاز به استخراج ویژگی برای پیش‌بینی شاخص بورس نیست و باید تمرکز بر روی انتخاب ویژگی‌های مناسب برای آن باشد.

جدول ۷. نتایج استخراج ویژگی با PCA در مدل SVR

۲۹ بعد ویژگی	درصد خطا	MAE	MSE	RMSE
مدل SVR	۳/۸۹	۰/۰۰۳۶	۰/۰۰۰۰۳۶	۰/۰۰۵۹

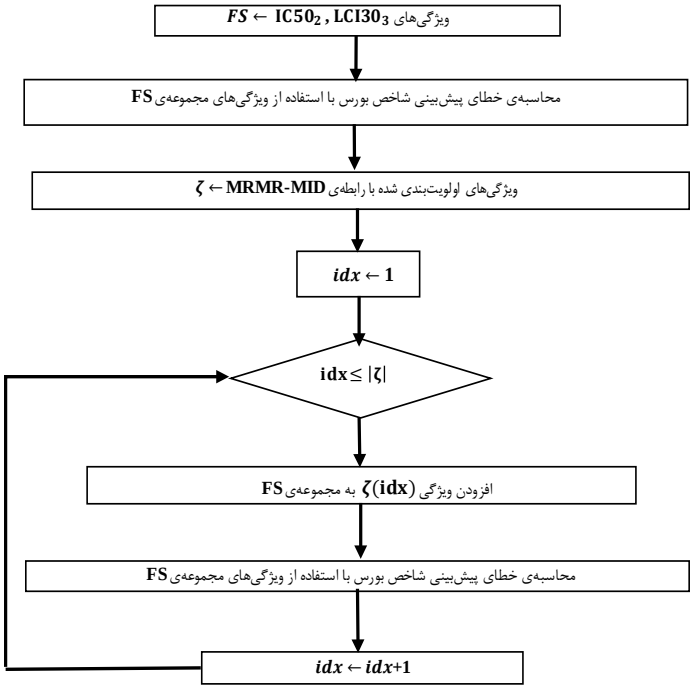


نمودار ۲: استفاده از دو تکنیک کاهش ابعاد برای پیش‌بینی شاخص بورس با بهره‌گیری از مدل SVR

الگوریتم پیشنهادی برای انتخاب ویژگی‌های مناسب برای مدل پیش‌بینی شاخص روزانه بورس

در الگوریتم پیشنهادی برای انتخاب ویژگی‌های مناسب، تعدادی از ویژگی‌ها توسط محقق انتخاب می‌شوند. براساس الگوی تغییر مقدار خطا با افزایش تعداد ویژگی‌ها در روش انتخاب ویژگی با الگوریتم MID، که در نمودار ۱ نمایش داده شده است، مشاهده می‌شود که در مدل SVR، دو ویژگی بازده شاخص ۳۰ شرکت بزرگ در۳ روز گذشته (ویژگی شماره ۹) و بازده شاخص ۵۰ شرکت فعال‌بورس در ۲ روز گذشته (ویژگی شماره ۱۵)، باعث کاهش قابل توجه خطای پیش‌بینی شده‌اند. بر این اساس، تصمیم بر آن شد که

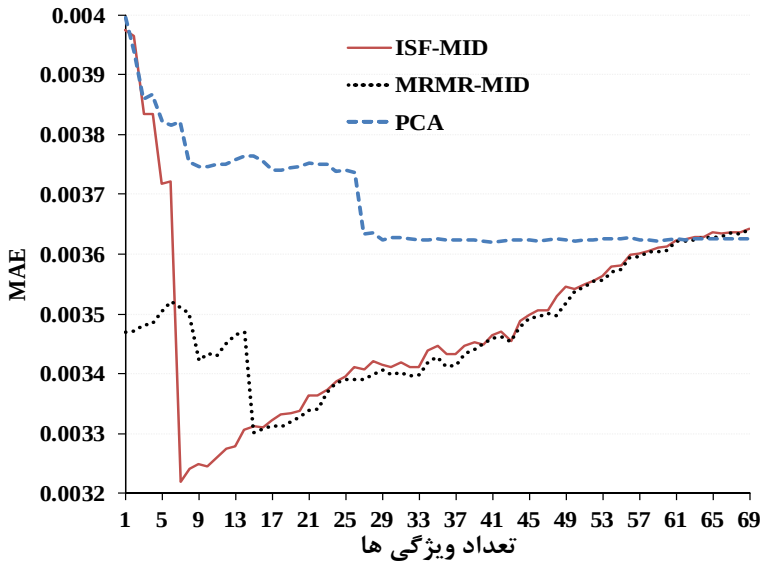
برای اولویت‌بندی ویژگی‌ها ابتدا، این دو ویژگی را به‌عنوان ویژگی‌های معرفی شده، انتخاب کرده و سپس به اولویت‌بندی سایر ویژگی‌ها با الگوریتم MID پرداخته شود. ویژگی‌ها با اولویت جدیدی که برای آن‌ها به‌دست آمده است، به مدل SVR به‌منظور پیش‌بینی شاخص بورس به ترتیب داده می‌شوند. فلوچارت الگوریتم پیشنهادی تحت عنوان ISF_MID[1] در شکل ۱ نمایش داده شده است. شایان ذکر است که در این ارزیابی نیز از روش اعتبارسنجی k-fold (K=10) استفاده می‌شود.



شکل۱. فلوچارت الگوریتم پیشنهادی برای انتخاب ویژگی‌های مناسب به‌منظور پیش‌بینی شاخص بورس

همچنین نتیجه این محاسبات که اثر افزایش تعداد ویژگی‌ها با اولویت مشخص‌شده را در دقت مدل SVR نشان می‌دهد، درنمودار ۳ نمایش داده شده است. در این نمودار، برای سهولت مقایسه استخراج ویژگی‌ها با الگوریتم PCA، اولویت‌بندی ویژگی‌ها با الگوریتم MID، و الگوریتم پیشنهادی ISF_MID، بر دقت عملکرد مدل SVR در پیش‌بینی روزانه شاخص کل بورس اوراق بهادار تهران نشان داده شده است.

1 .Initial Selected Features _Mutual Information Difference



نمودار۳. مقایسه استفاده از PCA، MID، و ISF-MID در انتخاب ویژگی‌های مناسب در مدل SVR به‌منظور پیش‌بینی روزانه شاخص کل بورس اوراق بهادار تهران

با مقایسه نمودارهای موجود می‌توان به این نتیجه رسید که مدل SVR با ویژگی‌های انتخابی توسط الگوریتم ISF_MID، تنها با ۷ ویژگی، بهترین نتیجه و دقت را ارائه کرده است. در صورتی‌که مدل SVR با ویژگی‌های انتخابی توسط الگوریتم MID، با ۱۵ ویژگی، و با ویژگی‌های استخراج شده توسط الگوریتم PCA با ۲۹ بعد، بیشترین دقت را دارد. نتایج استفاده از ویژگی‌های انتخابی توسط الگوریتم ISF_MID، در جدول۸ نشان داده شده است. با توجه به نتایج ارائه شده در این جدول می‌توان به‌این نتیجه رسید که؛ مدل SVR با الگوریتم ISF_MID، بهترین عملکرد را دارد. همان‌طور که ملاحظه می‌شود، مقدار MAE مدل SVR با ویژگی‌های انتخابی توسط الگوریتم ISF_MID، با ۷ ویژگی ابتدایی، برابر با ۰/۰۰۳ است.

جدول ۸. نتایج انتخاب ویژگی با ISF_MID در مدل SVR

۷ ویژگی	درصد خطا	MAE	MSE	RMSE
مدل SVR	۳/۴۶	۰/۰۰۳۲	۰/۰۰۰۰۲۸	۰/۰۰۵۲

این ویژگی‌ها به ترتیب عبارتند از: بازده شاخص ۵۰ شرکت فعال‌بورس در ۲ روز گذشته، بازده شاخص ۳۰ شرکت بزرگ در۳ روز گذشته، بازده شاخص کل در۱روز گذشته، تغییر نسبی حجم معاملات در ۳ روز

گذشته، تفاوت نسبی درصد بازده شاخص در ۱۰روز گذشته، تغییر قیمت سکه در ۲ روز گذشته، و تفاوت نسبی درصد بازده شاخص در ۵ روز گذشته.

نتیجه گیری و بحث

در این پژوهش سعی شده است با رویکردی کاملا جدید و منحصر به‌فرد، ویژگی‌های مناسب برای مدل پیش‌بینی روزانه شاخص کل بورس اوراق بهادار تهران شناسایی شوند. در این راستا به‌منظور انتخاب ویژگی‌های مناسب برای ورودی مدل SVR، الگوریتمی به نام ISF_MID، پیشنهاد شد که باعث افزایش دقت مدل ارائه شده به میزان ۹۶/۶۴ درصد گردید. برای شناسایی این الگوریتم ابتدا از هر دو تکنیک کاهش ابعاد (انتخاب و استخراج ویژگی) برای داده‌کاوی و پیش‌پردازش داده‌های ورودی به مدل SVR استفاده گردید، تا تاثیر هر دو روش کاهش ابعاد در دقت عملکرد مدل SVR در پیش‌بینی شاخص روزانه بورس اوراق بهادار تهران شناسایی گردد. برای استخراج ویژگی‌ها از PCA، و برای انتخاب ویژگی‌ها از دو روش تخمین mRMR با نام‌های MID و FCD، استفاده گردید. با مشخص شدن الویت ویژگی‌های انتخاب شده توسط MID و FCD، به مقایسه میزان دقت مدل SVR در استفاده از هر دو الگوریتم انتخاب ویژگی پرداخته شد، و براساس نتایج، MID به‌عنوان الگوریتم بهتر در انتخاب ویژگی تأیید شد. سپس با مقایسه الگوریتم MID با الگوریتم PCA، عملکرد بهتر استفاده از روش انتخاب ویژگی‌ها نسبت به استخراج ویژگی-ها در انتخاب متغیرهای ورودی مدل SVR مشاهده شد. بر این اساس، نتیجه گرفته شد که ضرورتی به استخراج ویژگی برای پیش‌بینی شاخص بورس نمی‌باشد و باید تمرکز بر روی انتخاب ویژگی‌های مناسب برای آن باشد. در نهایت با توجه به نتایج استفاده از MID در الویت‌بندی ویژگی‌های موثر بر پیش‌بینی شاخص بورس، الگوریتم ISF_MID، پیشنهاد گردید. با استفاده از این الگوریتم تنها با ۷ ویژگی بهترین نتیجه و دقت را می‌توان با مدل SVR در پیش‌بینی روزانه شاخص کل بورس اوراق بهادار تهران، بدست آورد.

به طور خلاصه؛ اقداماتی که در این پژوهش صورت گرفت که باعث تمایز آن با سایر پژوهش‌ها در این حوزه می‌شود، به شرح زیر می‌باشند:

- جمع‌آوری یک مجموعه کامل از ویژگی‌های تأثیرگذار بر شاخص بورس براساس مبانی تئوریک و بررسی پژوهش‌های پیشین؛ به‌طوری‌که ۶۹ ویژگی از ۱۶ شاخص مورد بررسی قرار گرفت.
- این مقاله، از روش اعتبارسنجی متقابل k-fold، برای انتخاب مجموعه‌های آموزش و آزمایش بهره برده‌است. این روش باعث می‌شود همه داده ها یک بار برای آموزش و یک بار برای آزمایش به‌کار روند، که این امر منجر به افزایش دقت و مفید واقع شدن مدل مورد نظر در عمل می‌گردد.
- مطالعات موجود در حوزه پیش‌پردازش داده‌ها به‌منظور انتخاب ویژگی‌های مناسب برای ورودی مدل پیش‌بینی، اتکا به استفاده از یک نوع تکنیک کاهش ابعاد (انتخاب یا استخراج ویژگی) را نشان می‌دهند. این امر ممکن است برخی از مفروضات مهم در مورد عملکرد رگرسیون اصلی متصل به متغیرهای ورودی و خروجی را نادیده بگیرد. بنابراین در این مطالعه، همزمان از دو

شیوه متفاوت کاهش ابعاد استفاده می‌شود. به‌طوری‌که برای الویت‌بندی و انتخاب ویژگی‌های مناسب از mRMR-MID وmRMR-FCD استفاده می‌شود، و برای استخراج ویژگی‌ها PCA به‌کار گرفته می‌شود. بنابراین برای نخستین‌بار این پژوهش یک فرآیند داده‌کاوی جامع را برای پیش‌بینی شاخص روزانه بورس انجام می‌دهد.

- برخلاف پژوهش‌های صورت گرفته در این حوزه که از مقدار همبستگی ویژگی‌های ورودی نسبت به خروجی مدل برای انتخاب ویژگی‌های مؤثر استفاده می‌شود، در این پژوهش برای انتخاب ویژگی‌های مناسب از mRMR استفاده گردید. این روش آماری؛ ویژگی‌های مؤثر با توجه به بیشینه‌سازی معیار وابستگی آماری مجموعه ویژگی‌ها با ویژگی هدف، و کمینه کردن اطلاعات متقابل در بین مجموعه ویژگی‌های انتخابی، گزینش می‌شوند.
- الگوریتمی تحت عنوان ISF_MID برای انتخاب ویژگی‌های مناسب مدل پیش‌بینی شاخص روزانه بورس اوراق بهادار تهران پیشنهاد گردید که با توجه به این الگوریتم می‌توان با تعداد معدودی ویژگی (۷ ویژگی) به پیش‌بینی با بیشترین دقت اقدام نمود، به‌طوری‌که میزان درصد خطا در مدل پیشنهادی به ۳/۴۶ رسیده است.

با توجه به نتایج حاصل از پژوهش، پیشنهاد می‌شود به‌منظور بررسی عملکرد الگوریتم پیشنهادی ISF_MID، مجموعه داده‌های مورد بررسی را گسترش داد و/ یا از دیگر تکنیک‌های هوشمند به عنوان مدل پیش‌بنی استفاده نمود، و مجدد عملکرد الگوریتم پیشنهادی در این پژوهش را مورد آزمون قرار داد. همچنین برای تکمیل آن توصیه می‌شود در پژوهش‌های بعدی از داده‌های بازار بورس دیگر کشورها استفاده نموده و این الگوریتم در این بازارها مورد مطالعه قرار گیرد.

در پایان خاطر نشان می‌کنیم که از نتایج این مطالعه پژوهشگران حوزه مدل‌سازی و سرمایه‌گذاران اعم از حقیقی و حقوقی می‌توانند استفاده کنند.

ملاحظات اخلاقی

حامی مالی: مقاله حامی مالی ندارد.
مشارکت نویسندگان: تمام نویسندگان در آماده‌سازی مقاله مشارکت داشته‌اند.
تعارض منافع: بنا بر اظهار نویسندگان در این مقاله هیچ‌گونه تعارض منافعی وجود ندارد.

منابع

Bajalan, S., Fallahpour, S., Dana, N. (2017). “Predicting stock price trends using a modified support vector machine with hybrid feature selection”. *Financial Management Perspective*, 7(17), 69-86. (In Persian).

Bustos, O. Pomares-Quimbaya, A. (2020). “Stock Market Movement Forecast: A Systematic Review”, *Expert Systems with Applications, Volume* 156, 15 October, 113464.

Cavalcante, R. C., Brasileiro, R. C. , Souza V. L.F., Nobrega, J. P. & Oliveira A. L.I. (2016). “Computational Intelligence and Financial Markets: A Survey and Future Directions”, *Expert Systems with Applications*. 55.194-211.

Ding, C. and H. Peng (2005). “Minimum redundancy feature selection from microarray gene expression data”. *Journal of bioinformatics and computational biology*. 3(2), 185-205.

Guo-Qiang, X. (2011). “The optimization of share price prediction model based on 1712 support vector machine”. In *International conference on control, automation and* 1713 *systems engineering* (pp. 1–4). IEEE.

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). “Literature review: Machine learning techniques applied to financial market prediction”. *Expert Systems with Applications*. Volume 124, 15 June. 226-251.

Huang, C.-F. (2012). “A hybrid stock selection model using genetic algorithms and support vector regression”. *Applied Soft Computing*, 12 (2), 807–818.

Kara, Y. , Boyacioglu, M. A. , & Baykan, O. K. (2011). “Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul stock exchange”. *Expert Systems with Applications*, 38 (5), 5311–5319.

Kumar, Deepak. Sarangi, Pradeepta Kumar & Verma, Rajit. (2021). “A systematic review of stock market prediction using machine learning and statistical techniques”, *Materials Today*: Proceedings.

Lee, Ming.Chi (2009). “Using support vector machine with a hybrid feature selection method to the stock trend prediction”. *Expert Systems with Applications.*Volume 36. Issue 8, 10896-10904.

Lui, Y., and Zheng, Y.F. (2006). “FS_SFS: A novel feature selection method for support vector machines”. *Pattern Recognition*. Volume 39, Issue 7, July 2006, Pages 1333-1345.

Mandal. M and Mukhopadhyay. A. (2013). “An improved minimum redundancy maximum relevance approach for feature selection in gene expression data”. *Procedia Technol*.10, 20–27.

Mansourfar, Gholamreza. Ghayour, Farzad, Khaleghparast Athari, Shabnam. (2015). “Predicting the Industry Index Volatility of Companies Listed in Tehran Stock Exchange, Emphasizing on Corporate Financial Variables Using Support Vector

Machine". *Journal of Empirical Studies in Financial Accounting*, Volume: 12 Issue: 46. (In Persian).

Monajemi, Amirhassan Ebrazi, Medi & Rayati, Alireza. (2009). "Stock price prediction in Tehran stock exchange using artificial neural network". *Journal of financial economy, 6(3),* 1-26. (In Persian).

Nevasalmi, Lauri. (2020). "Forecasting multinomial stock returns using machine learning methods". The Journal of Finance and Data Science, *Volume 6*, 86-106.

Nguyen, Duc-Hien, Le Manh-Thanh. (2014). "A two-stage architecture for stock price forecasting by combining SOM and fuzzy-SVM", *International Journal of Computer Science and Information Security (IJCSIS),* Vol. 12, No. 8, August.

Ni, L.P., Ni, Zh. W., & Gao, Y.Zh. (2011). Stock trend prediction based on fractal feature selection and support vector machine. *Expert Systems with Applications*, 38(5): 5569-5576.

Ou, P., & Wang, H. (2009). "Prediction of stock market index movement by ten data mining techniques". *Modern Applied Science*, 3, P28.

Patel, J., Shah, S., Thakkar, P., and Kotecha, K. (2015). "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques". *Expert Systems with Applications*, 42(1):259–268.

Pearson, K. (1901). "On lines and planes of closest fit to systems of points in space". *Philosophical Magazine*, 2(6), 559–572.

Perez-Rodriguez, J. V., S. Torrab and J. Andrada-Felixa (2004). "STAR and ANN models: Forecasting performance on the Spanish Ibex-35 stock index". Journal of Empirical Finance. 12(3), 490–509.

Raee, R., Nikahd, A., Habibi, M. (2017). "The Index Prediction of Tehran Stock Exchange by Combining the Principal Components Analysis, Support Vector Regression and Particle Swarm Optimization". *Financial Management Strategy*, 4(4), 1-23. (In Persian).

Rafiuzzaman, M. (2014). "Forecasting Chaotic Stock Market Data using Time Series Data Mining". *International Journal of Computer Applications*. 101(10), 27–34.

Singh, R. and Srivastava, S. (2017). "Stock prediction using deep learning". *Multimedia Tools and Applications*, 76(18):18569–18584.

Ul Haq, Anwar. Zeb, Adnan. Lei, Zhenfeng & Zhang, Defu. (2021). "Forecasting daily stock trend using multi-filter feature selection and deep learning", *Expert Systems with Applications*, 168 (2021) 114444

Wanga, Diya & Zhao, Yixi (2020) .*"*Using News to Predicton Investor Sentiment: Based on SVM Model", *Procedia Computer Science.* Volume 174 .191–199

Wei, Z. (2012). *A svm approach in forecasting the moving direction Chinese stock indices*, Department of industrial and systems engineering, Thesis of Master of Sciences, Lehigh University.

Yuan, Y. (2013). “Forecasting the movement direction of exchange rate with polynomial smooth support vector machine”. *Mathematical and Computer Modelling, 57* (3), 932–944.

Zhang, X., Hu, Y., Xie, K., Wang, S., Ngai, E. W. T., & Liu, M. (2014). “A causal feature selection algorithm for stock prediction modeling”. *Neurocomputing*, 142. 48-59.

Zhong, X., & Enke, D. (2017). “Forecasting daily stock market return using dimensionality reduction”. *Expert Systems with Applications*, 67, 126–13.