راهبرد مدیریت مالی
سال چهارم، شماره پانزدهم
زمستان ۱۳۹۵
صص ۲۳-۱

دانشگاه الزهرا (س)
دانشکده علوم اجتماعی و اقتصادی
تاریخ دریافت: ۱۳۹۵/۰۶/۲۴
تاریخ تصویب: ۱۳۹۵/۰۸/۱۸

پیش بینی شاخص بورس اوراق بهادار تهران با ترکیب روش‌های آنالیز مولفه‌های اصلی، رگرسیون بردارپشتیبان و حرکت تجمعی ذرات[۱]

رضا راعی[۲] علی نیک عهد قصیرائی[۳] و مصطفی حبیبی[٤]

چکیده

پیش‌بینی نوسان‌های آینده شاخص سهام می‌تواند اطلاعاتی در مورد روند آینده بازار سرمایه فراهم نماید. در این پژوهش، به منظور افزایش دقت پیش‌بینی شاخص بورس اوراق بهادار تهران، ترکیبی از روش‌های آماری و هوش مصنوعی به کار رفته است. مدل اصلی پیش‌بینی در این پژوهش، رگرسیون بردار پشتیبان بهینه شده به وسیله الگوریتم حرکت تجمعی ذرات می‌باشد. در برازش مدل رگرسیون بردار پشتیبان، سه پارامتر توضیحی وجود دارد که باید ترکیبی از این سه پارامتر توسط کاربر و به صورت آزمایش و خطا انتخاب شود تا دقت مدل را به بیشترین حد خود برساند. با توجه به زمان‌بر بودن و کارایی پایین انتخاب پارامتر توسط کاربر، برای انتخاب ترکیب بهینه پارامترهای مدل رگرسیون بردار پشتیبان، از روش بهینه‌سازی حرکت تجمعی ذرات استفاده شده است که الگوریتمی قوی در حوزه بهینه‌سازی می‌باشد. با توجه به حجم زیاد داده‌های ورودی به مدل برای کاهش زمان یادگیری و افزایش دقت پیش‌بینی، با استفاده از روش آنالیز مولفه‌های اصلی، پیش‌پردازش روی متغیرهای ورودی صورت گرفته و به مولفه‌های اصلی تبدیل شده است. نتایج بدست آمده نشان داد که پیش‌پردازش روی دادها، خطای پیش‌بینی مدل را به طور قابل ملاحظه‌ای کاهش داده است.

واژه‌های کلیدی: شاخص بورس، آنالیز مولفه‌های اصلی، رگرسیون بردار پشتیبان، بهینه‌سازی حرکت تجمعی ذرات، پیش‌بینی
طبقه بندی موضوعی: G10,G19,G17,C02

---

۱. کد DOI مقاله: 10.22051/jfm.2017.12410.1175

۲. استاد دانشگاه تهران، دانشکده مدیریت، گروه مدیریت مالی و بیمه، Email: raei@ut.ac.ir

۳. کارشناس ارشد مدیریت مالی، دانشگاه تهران، Email: h.nikahd@gmail.com

٤. دانشجوی کارشناسی ارشد مدیریت مالی، دانشگاه تهران، نویسنده مسئول، Email: mostafahabibi_68@yahoo.com

مقدمه

بورس اوراق بهادار، مکانی است که در آن پس‌اندازهای راکد جمع‌آوری شده و در تامین مالی پروژه‌های سرمایه‌گذاری بلند مدت استفاده می‌شود. افراد دارای پس‌اندازهای راکد، بازدهی سرمایه‌گذاری در بورس را با سایرگزینه‌های سرمایه گذاری(سرمایه‌گذاری در بخش مسکن، سرمایه‌گذاری در بانک، طلا، تولید مستقیم و ...ه) مقایسه کرده و با توجه به ریسک و بازده هر یک از سرمایه‌گذاری‌های ممکن، تصمیم گیری می‌کنند. تصمیم‌گیری درست، نیازمند اطلاعات است. این اطلاعات همیشه به‌طور کامل در دسترس نیست. بنابراین برای تصمیم‌گیری نیاز به پیش‌بینی داریم. مسئله پیش‌بینی شاخص سهام از دیرباز مورد توجه پژوهشگران بازار سرمایه قرار داشته و بدین منظور از مدل‌های خطی و غیرخطی زیادی استفاده شده است. اگر پیش‌بینی شاخص سهام به درستی، اطلاعات مربوط به روند آتی این متغیر را منعکس کند، می‌توان ازآن به عنوان یک متغیر پیشرو برای پیش‌بینی نوسان فعالیت‌های اقتصادی استفاده‌کرد. اما از انجایی که متغیرهای زیاد اثرگذار بر روی شاخص بازار اوراق بهادار را می‌توان شناسایی کرد و همچنین به دلیل این که سری زمانی شاخص از یک الگوی خطی پیروی نمی‌کند، برای کاهش خطای پیش‌بینی در این پژوهش، ترکیبی از روش‌های آماری و هوش مصنوعی استفاده شده است. در این پژوهش ابتدا از روش تجزیه و تحلیل مولفه‌های اساسی(PCA)[1]برای پالایش اولیه داده‌ها استفاده شده است. سپس با استفاده از رگرسیون بردار پشتیبان(SVR)[2]که نوع خاصی از ماشین‌های بردار پشتیبان(SVM)[3] می‌باشد، به پیش‌بینی شاخص اقدام شده است. در نهایت با استفاده از روش بهینه‌سازی حرکت تجمعی ذرات(PSO)[4]که یک روش بهینه‌سازی تکاملی می‌باشد، پارامترهای مدلSVR طوری انتخاب شده که خطای پیش‌بینی به کمترین حد خود برسد.

مبانی نظری و مروری بر پیشینه پژوهش

تحلیل مولفه اصلی تبیین ساختار واریانس-کوواریانس با چند ترکیب خطی از متغیرهای اصلی سر و کار دارد. اهداف کلی آن عبارتند از ۱- کاهش حجم داده‌ها و ۲- تعبیر و تفسیر آن‌ها.

---

1. Principal component analyses
2. Support Vector Regression
3. Support Vector Machine
4. Particle Swarm Optimization

اگر چه برای مطالعه تغییرپذیری کل سیستم، p مولفه لازم است، ولی بیشتر این تغییرپذیری را می‌توان با تعداد کمتر برای مثال k مولفه اصلی بیان نمود. در این صورت میزان اطلاع موجود در k مولفه (تقریبا) مانند میزان اطلاع در p متغیر اولیه است. بنابراین k مولفه اصلی را می‌توان به جای p متغیر اولیه به کار برد و مجموعه داده‌های اولیه شامل n اندازه روی p متغیر را به مجموعه‌ای از داده‌های شامل n اندازه درمورد k مولفه اصلی کاهش داد.

تحلیل مولفه‌های اصلی وسیله‌ای برای رسیدن به هدف هستند تا این‌که خودشان هدف باشند. زیرا آن‌ها اغلب به عنوان مراحل میانی در وضعیت‌های بزرگ‌تر به کار می‌آیند. برای مثال، مولفه‌های اصلی می‌توانند ورودی‌های یک رگرسیون چندگانه یا تحلیل خوشه‌ای باشند(جانسون، ۱۳۷۸).

مولفه‌های اصلی از نظر جبری، ترکیبات خطی ویژه p متغیر تصادفی $X_1, X_2, \ldots, X_p$ است. این ترکیبات خطی از نظر هندسی، انتخاب یک دستگاه مختصات جدید را نشان می‌دهد که از دوران دستگاه اولیه با $X_1, X_2, \ldots, X_p$ به عنوان محورهای مختصات به‌دست می‌آید. محورهای جدید، جهت‌ها را با بیشترین تغییرپذیری نشان می‌دهد و به بیان ساده‌تر، ساختمان کوواریانس‌ها را فراهم می‌کند.

چنان‌که ملاحظه خواهیم نمود، مولفه‌های اصلی تنها به ماتریس کوواریانس (یا ماتریس همبستگی) $X_1, X_2, \ldots, X_p$ مربوط می‌شود. برای گسترش آن‌ها، گسترش نرمال چند متغیری لازم نیست. از سوی دیگر، مولفه‌های اصلی جامعه‌های نرمال چند متغیری، تعابیر مفیدی بر حسب بیضوی‌های چگالی ثابت دارد. علاوه بر این، در جامعه نرمال چند متغیری، استنباط‌هایی را از مولفه‌های نمونه می‌توان به عمل آورد.

فرض کنید بردار تصادفی $X' = [X_1, X_2, \ldots, X_p]$ دارای ماتریس کوواریانس با مقادیر ویژه $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ است.

ترکیبات خطی زیر را در نظر می‌گیریم:

$$\begin{aligned} Y_1 &= e_1'X = e_{11}X_1 + e_{21}X_2 + \cdots + e_{p1}X_p \\ Y_2 &= e_2'X = e_{12}X_1 + e_{22}X_2 + \cdots + e_{p2}X_p \\ Y_p &= e_p'X = e_{1p}X_1 + e_{2p}X_2 + \cdots + e_{pp}X_p \end{aligned} \qquad (۱)$$

در این صورت داریم:

$$Var(Y_i) = e_i'\Sigma e_i \quad i = 1,2,\ldots,p \qquad (۲)$$

$$Cov(Y_i, Y_k) = e_i'\Sigma e_k \qquad i,k = 1,2,\ldots,p \qquad (۳)$$

مولفه‌های اصلی آن، ترکیبات خطی ناهمبسته $Y_1, Y_2, \ldots, Y_p$ هستند که واریانس‌های آن‌ها تا حد ممکن، بزرگ می‌باشد.

نخستین مولفه اصلی، یک ترکیب خطی با واریانس ماکزیمم است. یعنی $Var(Y_i) = e_1' \Sigma e_1$ را ماکزیمم می‌کند. واریانس دومین مولفه اصلی، کمتر از مولفه نخست می‌باشد. واریانس مولفه ها کاهش می‌یابد تا این‌که مولفه p ام کمترین واریانس را دارا دارد.

از تجزیه و تحلیل مولفه‌های اصلی نتایج زیر حاصل می‌شود:

نتیجه ۱- فرض کنید ماتریس کوواریانس بردار تصادفی $X' = \left[X_1, X_2, \ldots, X_p\right]$ باشد. فرض کنید دارای زوج مقدار ویژه-بردار ویژه $(\lambda_1, e_1), (\lambda_2, e_2), \ldots, (\lambda_p, e_p)$ باشد که $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$، مولفه اصلی iام با

$$\mathrm{Y_i = e_i'X = e_{1i}X_1 + e_{2I}X_2 + \cdots + e_{pi}X_p} \quad , \quad i = 1,2,\ldots,p \tag{۴}$$

داده شود. با این انتخاب‌ها داریم :

$$\mathrm{Var(Y_i) = e_i'\Sigma e_i = \lambda_i} \quad i = 1,2,\ldots,p \tag{۵}$$

$$Cov\left(Y_i {\scriptstyle و} Y_k\right) = e_i'\Sigma e_k = 0 \qquad i \neq k \tag{۶}$$

در صورتی که بعضی از λ_i ها برابر باشند ، انتخاب‌های بردار ضرایب مربوط e_i و در نتیجه Y_i یکتا نخواهد بود.

نتیجه۲- فرض کنید فرض کنید ماتریس کوواریانس بردار تصادفی $X' = \left[X_1, X_2, \ldots, X_p\right]$ باشد. فرض کنید دارای زوج مقدار ویژه-بردار ویژه $(\lambda_1, e_1), (\lambda_2, e_2), \ldots, (\lambda_p, e_p)$ باشد که $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ باشد. همچنین فرض کنید $Y_1 = e_1'X, Y_2 = e_2'X, \ldots, Y_p = e_p'X$ مولفه‌های اصلی باشند که در آن صورت داریم:

$$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{PP} = \sum_{i=1}^{P} Var(X_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p \tag{۷}$$

و در نتیجه نسبت واریانس کل مربوط به مولفه اصلی k ام عبارت است از:

$$\begin{pmatrix}\text{سهم کل واریانس}\\ \text{جامعه مربوط}\\ \text{به مولفه اصلی}\\ k\text{ ام}\end{pmatrix} = \frac{\lambda_k}{\lambda_1+\lambda_2+\cdots+\lambda_p} \qquad k=1,2,\ldots,p \qquad (۸)$$

نتیجه۳- اگر $Y_1 = e_1'X, Y_2 = e_2'X, \ldots, Y_p = e_p'X$ ، مولفه‌های اصلی به‌دست آمده از ماتریس کوواریانس باشد. آن گاه:

$$\rho_{Y_i,X_k} = \frac{e_{ki}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \qquad i,k=1,2,\ldots,p \qquad (۹)$$

ضرایب همبستگی بین مولفه‌های Y_i و متغیرهای X_k است. در این جا $(\lambda_1,e_1),(\lambda_2,e_2),\ldots,(\lambda_p,e_p)$ زوج‌های مقدار ویژه–بردار ویژه هستند(جانسون،۱۳۷۸).

روش ماشین بردار پشتیبان، یکی از روش‌های یادگیری ماشینی است که بر مبنای تئوری یادگیری آماری وپنیک[1] در دهه ۹۰ میلادی توسط وپنیک و همکارانش عرضه شد. در SVM، از اصول کمینه‌سازی ریسک ساختاری(SRM)[2] استفاده شده است، در حالی که سایر روش‌ها از اصول کمینه‌سازی ریسک تجربی (ERM)[3] بهره می‌برند (لیپ، ۲۰۰۵).

از ماشین بردار پشتیبان به طور کلی در مسائل طبقه‌بندی دو یا چند کلاسه و رگرسیون استفاده می‌شود. مانند بسیاری از روش‌های یادگیری ماشینی، در ماشین بردار پشتیبان نیز فرآیند ساخت مدل شامل دو مرحله آموزش و آزمایش می‌باشد. در انتهای فاز آموزش، قابلیت تعمیم‌یابی مدل آموزش داده شده با استفاده از داده‌های آزمایش ارزیابی می‌شود.

به طور خلاصه ساز و کار اصلی SVM در حل مساله رگرسیون به صورت زیر بیان می‌شود:

۱- ماشین بردار پشتیبان، تابع رگرسیون را با به کارگیری یک دسته تابع خطی تخمین می‌زند.

---

1. Vapnik
2. Structural Risk Minimization
3. Emprical Risk Minimization

۲- ماشین بردارپشتیبان، عملیات رگرسیون را با تابعی انجام می‌دهد که انحراف از مقدار واقعی در آن به میزان کمتر از مجاز است(تابع ضرر[1]).

۳- ماشین بردارپشتیبان، با کمینه‌کردن ریسک ساختاری، بهترین جواب را می‌دهد (سنچس، ۲۰۰۳).

در ماشین بردار پشتیبان برای حل مساله رگرسیون، یک تابع خطی به شکل $f(x) =< W.X > +b$، بر روی یک مجموعه شامل k نمونه مانند $\{(x_1, y_1),\ldots,(x_k, y_k)\} \in R^n, y \in R$، سعی در تخمین مقادیر خروجی بر مبنای مقادیر ورودی دارد. در آن رابطه، x بردار مقادیر ورودی و $(w,b)\ \epsilon R^n \times R$ پارامترهای کنترل کننده تابع fهستند. $< W.X >$ نشانگر ضرب داخلی می‌باشد. برای حل مساله رگرسیون، تابع ضرر وپنیک مورد استفاده قرار می‌گیرد که در آن کمترین خطا به میزان ε قابل صرف نظر کردن است. این تابع ضرر را می‌توان به صورت ذیل نمایش داد.

$$L_\varepsilon\left(X, Y, f(x)\right) = \begin{cases} |y - f(x)| - \varepsilon & |y - f(x)| \le \varepsilon \\ 0 & other\ wise \end{cases} \qquad (۱۰)$$

$L_\varepsilon(y)$ معرف تابع ضرر و خطای مجاز در تابع ضرر می‌باشد. پارامترهای کنترل کننده تابع رگرسیون بهینه با حل مساله بهینه‌سازی زیر به‌دست می‌آیند.

$$\text{minimize } \phi\left(W, \vartheta^*, \vartheta\right) = \frac{\|w\|^2}{2} + c\left(\sum \vartheta_j^* + \sum \vartheta\right)$$

$$subject\ to:$$

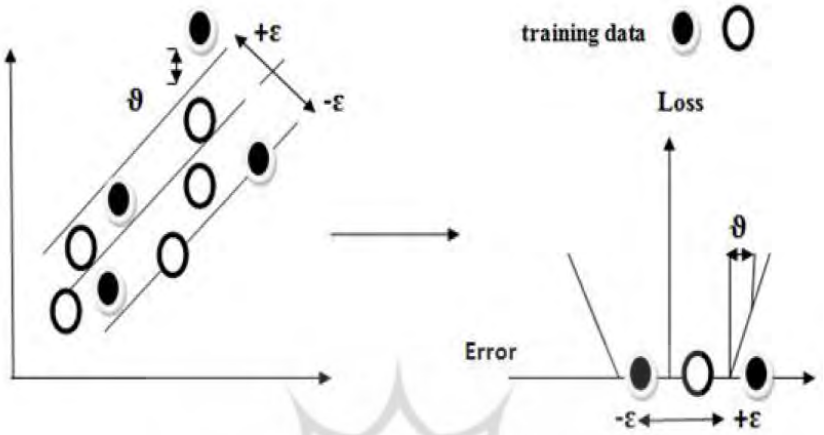
$$\left((W.X_j) + b\right) - y_j \le \varepsilon + \vartheta_j^* \qquad (۱۱)$$

$$y_j - \left((W.X_j) + b\right) \le \varepsilon + \vartheta_j$$

$$\vartheta_j, \vartheta_j^* \ge 0$$

1. Loss function

در رابطه ۱۱، و ϑ^* متغیرهای slack هستند. این متغیرها به همراه تابع ضرر در شکل ۱ نشان داده شده‌اند (گان 1998).



شکل ۱. تابع ضرر وپنیک و متغیرهای slack

برای حل مساله بهینه‌سازی فوق، به کمک تئوری لاگرانژ، تابع لاگرانژ به صورت زیر نوشته می‌شود.

$$L(a^*,a) = -\varepsilon\sum_{i=1}^{k}(a_i^* + a_i) + \sum_{i=1}^{k} y_i(a_i^* - a_i) - \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k}(a_i^*, a_i)(a_j^*, a_j)(x_i, x_j) \quad (۱۲)$$

با بیشینه شدن تابع فوق تحت قیدهای زیر ، مقادیر ضرایب a و a^* به‌دست می‌آیند. این ضرایب، ضرایب لاگرانژ نامیده می‌شوند.

$$\begin{cases} \sum a_i^* = \sum a_i \\ 0 \le a_i^* \le C \\ 0 \le a_i \le C \end{cases} \quad \text{for } i = 1, 2, 3, \dots, k \tag{۱۳}$$

مساله بهینه سازی فوق به کمک روش های برنامه ریزی درجه دو(QP) قابل حل می‌باشد. در نتیجه رسیدن به اکسترمم کلی نیز قطعی خواهد بود وخطر به دام افتادن در اکسترمم محلی وجود ندارد. داده‌هایی‌که ضرایب لاگرانژ متناظر با آنها غیرصفر باشد، به عنوان بردار پشتیبان شناخته می‌شوند. این داده‌ها از نظر هندسی دارای خطای پیش‌بینی بزرگتر از $\mp\varepsilon$ هستند. بنابراین بردارهای پشتیبان درون باند $\mp\varepsilon$ قرار نمی‌گیرند و مقدار ، تعداد بردارهای پشتیبان را کنترل می‌کند.

به کمک ضرایب لاگرانژ و بردارهای پشتیبان ، پارامترهای کنترل کننده پاسخ بهینه به صورت زیر محاسبه می‌شود.

$$W_0 = \sum (a_i^* - a_i) x_i \tag{۱۴}$$

$$b_0 = -\left(\frac{1}{2}\right) . W_0 . [x_r + x_s] \tag{۱۵}$$

$$f(x) = \sum (a_i^* - a_i)\,(x_i - x) + b_0 \tag{۱۶}$$

در رابطه فوق، x_s و x_r دو بردار پشتیبان هستند.

برای ساخت مدل ماشین بردار پشتیبان، پارامترهای c و توسط کاربر تعریف می‌شوند. پارامتر c یک پارامتر تنظیمی است و می‌تواند مقادیر صفر تا بی نهایت را بپذیرد. نقش این پارامتر، ایجاد تعادل میان کمینه کردن ریسک تجربی وبیشینه کردن قابلیت تعمیم یابی است. پارامتر نیز می‌تواند مقادیر صفر تا بی نهایت را بپذیرد. مقدار این پارامتر در وضعیت بردارهای پشتیبان و در نتیجه کارایی مدل بسیار موثر است.

مساله رگرسیون خطی در SVM به آسانی قابل گسترش به رگرسیون غیر خطی است. بدین منظور از توابع کرنل استفاده می‌شود. تاکنون کرنل‌های گوناگونی از جمله کرنل‌های چند جمله‌ای و پایه شعاعی (RBF) شناخته شده‌اند. بدین ترتیب در حالت رگرسیون غیرخطی SVM، پارامترهای کنترل کننده تابع بهینه با روابط زیر محاسبه می‌شوند (سنچس، ۲۰۰۳)

$$\mathrm{W.X} = \sum (a_i^* - a_i) K(x_i, x) \tag{۱۷}$$

$$\mathrm{b_0} = -\frac{1}{2}\sum (\mathrm{a_i^* - a_i})\left[\mathrm{K(x_r, x_i) + K(x_s, x_i)}\right] \tag{۱۹}$$

در این روابط $K(x_i, x)$ نشانگر تابع کرنل می‌باشد.

موجودات طبیعی گاهی به صورت یک توده رفتار می کنند. یکی از جریانهای اصلی در پژوهش زندگی مصنوعی، بررسی چگونگی رفتار موجودات طبیعی به صورت یک توده و پیاده سازی دوباره مدلِ توده ها در رایانه است.

یک روش جدید بهینه سازی با استفاده از همانند سازی رفتارِ گروهیِ موجودات طبیعی در اوایل دهه ۱۹۹۰ ابداع شد. ابرهارت و کندی(۱۹۹۵)، بهینه سازی ذرات انبوه (PSO) را بر اساس شبیه سازی از توده های پرندگان و دسته ماهی ها توسعه دادند. هر فرد، تجارب قبلی خود را در PSO مبادله می کند. PSO برای حل مسایل بهینه سازی غیرخطی با متغیرهای پیوسته ایجاد شده است. علاوه بر این، بر خلاف روش های تکاملی دیگر مانند الگوریتم ژنتیک، PSO می تواند با تنها یک برنامه کوچک پیاده‌سازی شود. این قابلیت PSO، یکی از مزیت‌های آن در مقایسه با دیگر تکنیک‌های بهینه سازی است. PSO روشی مبتنی بر تکنیک هایی تصادفی است که از آن می توان برای پیدا کردن مینیمم سراسری (غیر قطعی) مسایل برنامه ریزی غیرخطی استفاده کرد.

کندی و ابرهارت، PSO را از طریق شبیه سازی دسته پرندگان توسعه دادند. موقعیت هر عامل با s و همچنین سرعت آن با v نمایش داده می شود. اصلاح موقعیت عامل با استفاده از اطلاعات موقعیت و سرعت صورت می گیرد.

دسته پرندگان، یک تابع هدف خاص را بهینه سازی می کنند. هر عامل بهترین مقدار تجربه کرده را pbest و موقعیت فعلی را s می داند. این اطلاعات، تجربیات شخصی هر عامل است. علاوه بر

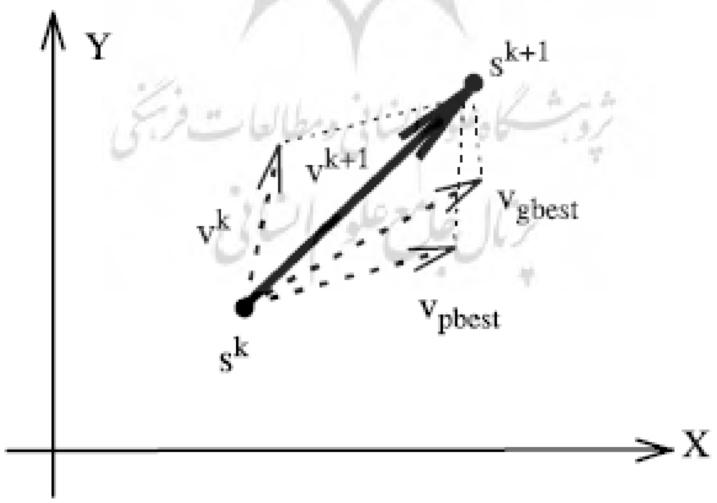
این، هر عامل، بهترین مقدار به دست آمده در گروه را gbest می داند. هر عامل تلاش می کند تا موقعیت خود را با استفاده از اطلاعات زیر تغییر دهد:

موقعیت فعلی s، سرعت فعلی v ، بهترین موقعیت شخصی pbest ، بهترین موقعیت گروهی gbest. سرعت هر عامل را می توان از معادله زیر به دست آورد:

$$ _i^{k+1} = wv_i^k + c_1 randd_1 \times \left(pbest_i - s_i^k\right) + c_2 rand_2 \times \left(gbest - s_i^k\right) \quad (۲۰)$$

که در آن، v_i^k سرعت عامل i در تکرار k ام، $pbest_i$ ، pbest نقطه i است و gbest ، gbest گروه است.

معنای طرف راست معادله می تواند به صورت زیر بیان شود . طرف راست معادله دارای سه جمله است. نخستین جمله، سرعت قبلی عامل است. جملات دوم و سوم برای تغییر سرعت عامل است. بدون جملات دوم و سوم، عامل پرواز در جهت قبلی خود ادامه خواهد داد تا به مرز برخورد کند. عامل سعی می کند ناحیه های جدید را جستجو کند و بنابراین، نخستین جمله با تنوع در روند جستجو متناظر است. به عبارت دیگر، بدون جمله نخست، سرعت پرواز عامل تنها با استفاده از موقعیت فعلی و بهترین موقعیت آن در گذشته تعیین می شود. عوامل سعی خواهند کرد با pbest ها و یا gbest همگرا شوند.

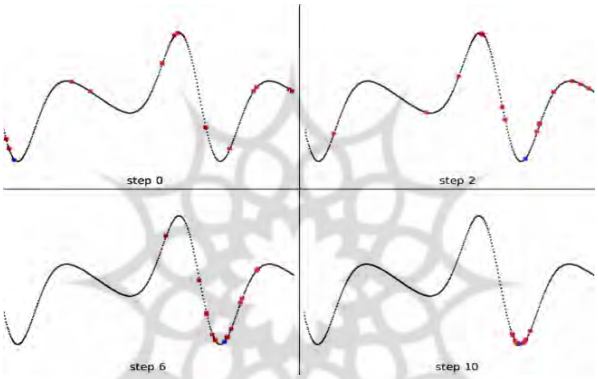


شکل ۲

موقعیت فعلی نقطه جستجو در فضای جواب را می توان با معادله زیر اصلاح کرد:

$$s_i^{k+1} = s_i^k + v_i^{k+1} \quad (۲۱)$$

هر عامل، موقعیت فعلی خود را با استفاده از ترکیب بردارهای نشان داده شده در شکل ۸ اصلاح می‌کند. در واقع، PSO، از چندین نقطه جستجو استفاده می کند و نقاط جستجو، تدریجی به نقطه بهینه با استفاده از pbest ها و gbest نزدیک می شوند.



شکل ۳. جست و جو با عوامل در فضای جواب با PSO

نوری و همکاران (۱۳۸۹) در مقاله‌ای تحت عنوان "پیش‌بینی ماهانه جریان آب با استفاده از ماشین بردار پشتیبان بر مبنای آنالیز مولفه اصلی" با هدف بررسی تاثیر انتخاب متغیرهای ورودی به کمک آنالیز مولفه اصلی، عملکرد مدل ماشین بردار پشتیبان را مورد بررسی قرار دادند. به این منظور ابتدا با استفاده از ۱۸ متغیر ورودی به مدل SVM، دبی جریان ماهانه پیش‌بینی شد. سپس با استفاده از آنالیز مولفه اصلی، تعداد متغیرهای ورودی به مدل ماشین بردار پشتیبان از ۱۸ متغیر به ۵ مولفه کاهش یافت. در نهایت با استفاده از آماره توسعه یافته توسط نویسندگان مقاله، عملکرد مدل‌های داده شده مورد ارزیابی قرار گرفت. این پژوهش نشان داد که پیش‌پردازش متغیرهای ورودی به SVM، بهبود عملکرد SVM را به همراه داشته است.

کیانی و همکاران(۱۳۸۷) در مقاله‌ای تحت عنوان" بررسی میزان دقت دسته‌بندی ماشین بردار پشتیبان در ارزیابی اعتباری بانکی" برای ارزیابی دقت دسته‌بندی، دو مجموعه داده اعتباری را با

استفاده از ماشین‌های بردار پشتیبان مورد تجزیه و تحلیل قرار دادند. با توجه به نتایج به‌دست آمده، دسته‌بندی کننده ماشین بردار پشتیبان در مقایسه با شبکه‌های عصبی، برنامه‌نویسی ژنتیکی و دسته‌بندی کننده درخت تصمیم با وجود در بر داشتن ویژگی‌های کمتر در ورودی به نتایج مشابهی دست پیدا می‌کنند. با تعیین میزان دقت نتایج به‌دست آمده می‌توان به این نتیجه رسید که ماشین بردارپشتیبان، یک روش جدید و قابل اطمینان در میان دیگر روش‌های داده‌کاوی است.

چی‌جی‌لو و همکاران (۲۰۱۳) در مقاله‌ای تحت عنوان "پیش‌بینی شاخص با مدل هیبریدی آنالیز مولفه‌های مستقل و رگرسیون بردارپشتیبان با بهینه‌سازی حرکت تجمعی ذرات" به پیش‌بینی شاخص بورس‌های چین، تایوان و هند پرداختند. آن‌ها در این مقاله، بیشترین، کم‌ترین و مقدار پایانی و مقدار باز شدن شاخص در روز جاری را به عنوان متغیر برای پیش‌بینی در نظر گرفتند. آن‌ها در پایان به این نتیجه رسیدند که مدل ترکیبی، شاخص را با خطای کمتری نسبت به مدل SVR پیش‌بینی می‌کند.

لی و همکاران (۲۰۱۴) در مقاله‌ای تحت عنوان " پیش‌بینی نفت خام با مدل‌های نوفه‌زدایی شده چندمقیاسه" اقدام به پیش‌بینی قیمت نفت خام نمودند. آن‌ها از مدل ARIMA برای پیش‌بینی قیمت نفت استفاده نمودند و چون این مدل، یک مدل خطی می‌باشد، برای پیش‌بینی بخش غیرخطی سری زمانی، مدل ماشین بردار پشتیبان را به کار بردند. در این پژوهش برای جلوگیری از تاثیر نوسان‌های قیمت نفت از نوفه‌زدایی موجک استفاده نمودند که نتایج حاکی از بهبود قدرت پیش‌بینی مدل بوده است.

چان چانگ و همکاران در سال ۲۰۱۵ در مقاله‌ای تحت عنوان "مدل فازی ترکیب شده با رگرسیون بردار پشتیبان برای پیش‌بینی معاملات سهام"، یک روش جدید برای شناسایی سیگنال‌های معاملاتی دادند. در این مقاله از یک مدل مبتنی بر قوانین فازی استفاده شده است که می‌توان سیگنال‌های معاملاتی را بر مبنای متغیرهای تکنیکال و رگرسیون بردارپشتیبان شناسایی نمود. این مدل با مدل‌های رگرسیون خطی معمولی و شبکه‌های عصبی مصنوعی مورد مقایسه قرار گرفته که نتایج نشان می‌دهد مدل پیشنهادی، بازده‌های بیشتری نسبت به مدل های رگرسیون خطی معمولی و شبکه‌های عصبی مصنوعی به دست می‌آورد.

فرضیه‌های پژوهش

در مورد فرضیه پژوهش باید بیان نمود که نظر به اثبات برتری دقت پیش‌بینی مدل رگرسیون بردار پشتیبان نسبت به سایر روش‌های پیش‌بینی در مطالعات گذشته، هدف اصلی این پژوهش، بهبود

پیش‌بینی رگرسیون بردار پشتیبان با استفاده از پیش‌پردازش داده‌ها به وسیله آنالیز مولفه‌های اصلی می‌باشد. بنابراین فرضیه پژوهش به صورت زیر بیان می‌شود:

پیش‌بینی مدل ترکیبی PCA-SVR-PSO نسبت به پیش‌بینی مدل SVR-PSO خطای کمتری دارد.

روش‌شناسی پژوهش

در پژوهش حاضر تلاش می‌شود مدلی از ترکیب روش‌های آماری و هوش مصنوعی برای پیش‌بینی شاخص بورس اوراق بهادار تهران داده شود.

با توجه به مطالب بخش مبانی نظری پژوهش، مولفه‌های اصلی وسیله‌ای برای رسیدن به هدف هستند تا این‌که خودشان هدف باشند. زیرا آن‌ها اغلب به عنوان مراحل میانی در وضعیت‌های بزرگ‌تر به کار می‌آیند. برای مثال مولفه‌های اصلی می‌توانند ورودی‌های یک رگرسیون چندگانه یا تحلیل خوشه‌ای باشند.

در این پژوهش برای پیش‌بینی شاخص بورس اوراق بهادار، مطابق با پژوهش لو و همکاران(۲۰۱۳)، از چهار متغیر شامل مقدار آغازین، بیشترین مقدار، کمترین مقدار و مقدار پایانی شاخص بورس اوراق بهادار استفاده شده است. به دلیل امکان وجود همبستگی و هم‌خطی میان داده‌ها که باعث ایجاد تورش در مقادیر پیش‌بینی می‌شود، با استفاده از روش آنالیز مولفه‌های اصلی، چهار متغیر ورودی به چهار مولفه اصلی پوشش دهنده کل پراکندگی داده‌ها تبدیل شده است. چهار مولفه اصلی استخراج شده از داده‌های ورودی به طور کامل، مستقل از یکدیگر می‌باشند که به عنوان ورودی مدل پیش‌بینی مورد استفاده قرار می‌گیرد.

به طور خلاصه سازوکار اصلی SVM در حل مساله رگرسیون به صورت زیر بیان می‌شود:

۱- ماشین بردار پشتیبان، تابع رگرسیون را با به کارگیری یک دسته تابع خطی تخمین می‌زند.

۲- ماشین بردار پشتیبان، عملیات رگرسیون را با تابعی انجام می‌دهد که انحراف از مقدار واقعی در آن به میزان کمتر از مجاز است(تابع ضرر).

۳- ماشین بردار پشتیبان، با کمینه‌کردن ریسک ساختاری، بهترین جواب را می‌دهد.

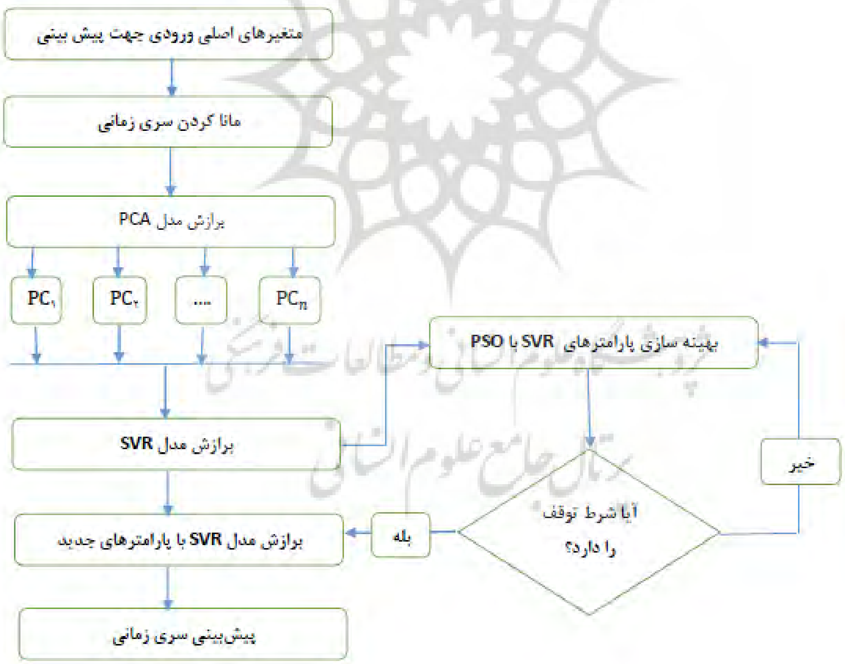
با توجه به توضیحات در باره مفهوم ریاضی مدل SVR، این مدل در نرم‌افزار MATLAB شبیه‌سازی شده است. سپس مولفه‌های اصلی خروجی از مدل PCA را به عنوان ورودی برای این مدل در نظر گرفته و مدل را با توجه به داده‌های آموزش برازش کرده و سپس با توجه به ورودی‌های

آزمایش، مقدار شاخص برای ۲۰ روز آینده را پیش‌بینی کرده و با مقادیر واقعی مقایسه می‌نماییم و با توجه به معیارهای ارزیابی MAPE و RMSE، دقت پیش‌بینی مدل را اندازه‌گیری می‌نماییم.

در مرحله قبل، مدل اصلی پیش‌بینی SVR را برازش نمودیم. برای ساخت مدل رگرسیون بردار پشتیبان، پارامترهای c و   و   توسط کاربر تعریف می‌شوند. پارامتر c، یک پارامتر تنظیمی است و می‌تواند مقادیر صفر تا بی‌نهایت را بپذیرد. نقش این پارامتر، ایجاد تعادل میان کمینه کردن ریسک تجربی و بیشینه کردن قابلیت تعمیم‌یابی است. پارامتر   نیز می‌تواند مقادیر صفر تا بی‌نهایت را بپذیرد. مقدار این پارامتر در وضعیت بردارهای پشتیبان و در نتیجه کارایی مدل بسیار موثر است.   نیز عرض کرنل با پایه شعاعی می‌باشد و مقدار آن توسط کاربر تعیین می‌شود.

چون کاربر ممکن است دقت لازم برای انتخاب این پارامترها را نداشته باشد، انتخاب این پارامترها را با مدل PSO انجام داده‌ایم تا خطای مدل SVR به کمترین حد خود برسد.

الگوریتم کلی مدل پیش‌بینی به شکل زیر می‌باشد:



شکل ٤. الگوریتم کلی مدل پیش‌بینی

تجزیه و تحلیل داده‌ها و آزمون فرضیه‌ها

در این قسمت، یافته‌های پژوهش و تجزیه و تحلیل آن‌ها آورده شده است. به منظور بررسی بهبود در قدرت پیش‌بینی مدل، داده‌های شاخص کل بورس اوراق بهادار و شاخص ۵۰ شرکت فعال‌تر از سال ۱۳۹۱ تا ۱۳۹۵، با استفاده از تکنیک پنجره غلتان[1] به پنج دوره زمانی برابر تقسیم شده و سپس برای هر دوره پیش‌بینی با مدل های SVR-PSO (استفاده از داده‌های معمولی برای پیش-بینی با رگرسیون بردار پشتیبان)و PCA-SVR-PSO(استفاده از مولفه‌های اصلی خروجی از روش آنالیز مولفه‌های اصلی، به عنوان وردی مدل رگرسیون بردار پشتیبان برای پیش‌بینی) در ۲۰ روز آینده انجام گرفته و سپس با مقادیر واقعی مقایسه و معیارهای میانگین مربع خطا و درصد قدر مطلق خطا استخراج شده است. سپس به منظور بررسی معنادار بودن تفاوت میانگین خطاهای دو مدل از آزمون‌های تی-استیودنت و دایبولد-ماریانو استفاده شده است.

با برازش هر یک از مدل‌های SVR-PSO و PCA-SVR-PSO، معیار ارزیابی عملکرد ریشه میانگین مربع خطا در جدول ۱ نشان داده شده است. همان‌طور که دیده می‌شود، مدل PCA-SVR-PSO نسبت به مدل SVR-PSO در تمامی دوره‌ها، عملکرد بهتری داشته است و از لحاظ میانگین ۵ دوره نیز خطای کمتری را نمایش می‌دهد. برای آزمون فرضیه پژوهش مبنی بر عملکرد بهتر مدل PCA-SVR-PSO نسبت به مدل SVR-PSO، معنادار بودن تفاوت میانگین این دو مدل بررسی شده است. معنی‌دار بودن تفاوت نشان می‌دهد که عملکرد مدل PCA-SVR-PSO به صورت معنی‌داری از مدل SVR-PSO بهتر است.

جدول ۱. ریشه میانگین مربع خطا

دوره	شاخص کل		شاخص ۵۰ شرکت فعالتر	
	SVR-PSO	PCA-SVR-PSO	SVR-PSO	PCA-SVR-PSO
۱	۱۴۸/۳۴	۱۰۸/۸۵	۱۸/۱۲	۹/۲۲
۲	۳۸۰/۴۴	۳۳۵/۴۶	۱۲/۴۰	۷/۷۶
۳	۴۱۳/۵۶	۳۴۴/۲۰	۳۰/۲۲	۲۰/۵۳
۴	۴۲۲/۳۳	۳۵۵/۲۴	۲۲/۴۵	۸/۴۴
۵	۴۴۸/۳۶	۳۴۵/۴۸	۱۵/۲۵	۵/۸۶
میانگین	۳۶۲/۶۱	۲۹۷/۸۵	۱۹/۶۹	۱۰/۳۶

1. Rolling window

با برازش هر یک از مدل‌های SVR-PSO و PCA-SVR-PSO، معیار ارزیابی عملکرد میانگین درصد قدرمطلق خطا در جدول ۲ نشان داده شده است. همان‌طور که دیده می‌شود، مدل PCA-SVR-PSO نسبت به مدل SVR-PSO در تعداد روزهای بیشتری، عملکرد بهتری داشته است و از لحاظ میانگین پیش‌بینی ۲۰ روزه نیز خطای کمتری را نمایش می‌دهد. برای آزمون فرضیه پژوهش مبنی بر عملکرد بهتر مدل PCA-SVR-PSO نسبت به مدل SVR-PSO، معنادار بودن تفاوت میانگین این دو مدل بررسی شده است. معنی‌دار بودن تفاوت نشان می‌دهد که عملکرد مدل PCA-SVR-PSO به صورت معنی‌داری از مدل SVR-PSO بهتر است.

جدول ۲. درصد قدرمطلق خطا

دوره	شاخص کل		شاخص ۵۰ شرکت فعالتر	
	SVR-PSO	**PCA-SVR-PSO**	**SVR-PSO**	**PCA-SVR-PSO**
۱	۰/۵۳۴	۰/۲۸۵	۰/۷۷۰	۰/۶۸۷
۲	۰/۸۵۲	۰/۴۳۲	۰/۷۰۴	۰/۶۶۱
۳	۰/۴۹۵	۰/۲۰۳	۰/۹۷۴	۰/۶۴۳
۴	۰/۷۴۱	۰/۵۴۵	۰/۸۵۴	۰/۴۹۸
۵	۱/۱۴۲	۰/۹۶۵	۱/۱۶۰	۰/۹۴۶
میانگین	۰/۷۵۳	۰/۴۸۶	۰/۸۹۲	۰/۶۸۷

در آزمون فرضیه نخست، عملکرد دو مدل PCA-SVR-PSO و SVR-PSO در پیش بینی شاخص کل مورد آزمایش قرار می گیرد. این فرضیه را از طریق هر دو معیار ارزیابی عملکرد میانگین قدرمطق درصد خطا و ریشه میانگین قدرمطلق خطا آزمون می کنیم.

$H_0 =$ بین دو جامعه میانگین قدرمطلق درصد خطای مدلهای $SVR - PSO$ و $PCA - SVR - PSO$ اختلاف معناداری وجود ندارد

$H_1 =$ بین دو جامعه میانگین قدرمطلق درصد خطای مدلهای $SVR - PSO$ و $PCA - SVR - PSO$ اختلاف معناداری وجود دارد

جدول ۳. آزمون مقایسه زوجی قدرمطلق درصد خطا-شاخص کل

آزمون مقایسه‌های زوجی قدرمطلق درصد خطای مدل های svr-pso و pca-svr-pso		
	SVR-PSO	PCA-SVR-PSO
میانگین	۰/۷۵۳	۰/۴۸۶
واریانس	۰/۰۶۹	۰/۰۸۹
تعداد مشاهده‌ها	۵	
درجه آزادی	۴	
آماره t	سطح خطای ۵ درصد	۶/۴۸۸
مقدار بحرانی آزمون یک طرفه		۲/۱۳۲
مقدار بحرانی آزمون دوطرفه		۲/۷۷۶
prob آزمون یک طرفه		۰/۰۰۱
آزمون دوطرفه prob		۰/۰۰۲۹۰۹۹۵۶

همان‌طور که آزمون مقایسه‌های زوجی بالا نشان می‌دهد، میانگین قدرمطلق درصد خطای مدل PCA-SVR-PSO از مدل SVR-PSO کم‌تر می‌باشد. با توجه به جدول فوق، آماره t محاسبه شده از مقدار بحرانی سطح خطای ۵ درصد بالاتر است. بنابراین در سطح خطای پنج درصد، فرضیه H_0 رد و فرضیه H_1 به اثبات می‌رسد که حاکی از تفاوت معنی‌دار قدر مطلق درصد خطای دو مدل است. در صورتی که آزمون فوق را به صورت یک طرفه انجام بدهیم، با توجه به جدول فوق، فرضیه H_0 مبنی بر بزرگ‌تر و یا مساوی بودن قدر مطلق درصد خطای مدل PCA-SVR-PSO نسبت به مدل SVR-PSO رد و فرضیه اصلی پژوهش مبنی بر بالا بودن دقت مدل PCA-SVR-PSO نسبت به مدل SVR-PSO در پیش‌بینی شاخص بورس تایید می‌شود.

$$\begin{cases} H_0 = \text{بین دو جامعه ریشه میانگین مربع خطای مدلهای } SVR-PSO \text{ و } PCA-SVR-PSO \text{ اختلاف معناداری وجود ندارد} \\ H_1 = \text{بین دو جامعه ریشه میانگین مربع خطای مدلهای } SVR-PSO \text{ و } PCA-SVR-PSO \text{ اختلاف معناداری وجود دارد} \end{cases}$$

جدول ٤. آزمون مقایسه زوجی ریشه میانگین مربع خطا-شاخص کل

آزمون مقایسه‌های زوجی ریشه میانگین مربع خطای مدل های svr-pso و pca-svr-pso		
	SVR-PSO	PCA-SVR-PSO
میانگین	۳۶۲/۶۱۰	۲۹۷/۸۵۰
واریانس	۱۴،۹۳۶/۲۴۶	۱۱،۲۱۱/۵۱۶
تعداد مشاهده‌ها	۵	
درجه آزادی	۴	
آماره t	سطح خطای ۵درصد	۶/۰۶۲
مقدار بحرانی آزمون یک طرفه		۲/۱۳۲
مقدار بحرانی آزمون دوطرفه		۲/۷۷۶
prob آزمون یک طرفه		۰/۰۰۲
آزمون دوطرفه prob		۰/۰۰۳۷۳۹۶۹

جدول فوق، نتایج آزمون مقایسه‌های زوجی ریشه میانگین مربع خطای مدل‌های PCA-SVR-PSO و SVR-PSO را نشان می‌دهد. نتایج جدول فوق حاکی از پایین‌تر بودن ریشه میانگین مربع خطای مدل PCA-SVR-PSO نسبت به مدل SVR-PSO می‌باشد. جدول فوق، رد فرضیه H_0 را در سطح خطای پنج درصد تایید می‌کند. بنابراین در سطح خطای پنج درصد، مدل PCA-SVR-PSO نسبت به مدل SVR-PSO با توجه به معیار ارزیابی عملکرد ریشه میانگین مربع خطا عملکرد بهتری در پیش بینی شاخص کل دارد.

همان‌گونه که گفته شد، علاوه بر آزمون مقایسه‌های زوجی برای بررسی معناداری تفاوت بین میانگین دو جامعه ریشه میانگین مجذور خطا و درصد قدرمطلق خطا، از دو آماره دایبولد-ماریانو و آماره هاروی-لیبورن-نیوبولد(آماره تعدیل شده دایبولد-ماریانو) نیز بهره گرفته شده است که نتایج آن به شرح جدول ۵ می‌باشد.

آزمون دایبولد-ماریانو برای معیارهای ارزیابی عملکرد محاسبه شده برای شاخص کل به صورت زیر می‌باشد:

جدول ۵. آزمون دایبولد-ماریانو

معیار ارزیابی عملکرد	آماره دایبولد-ماریانو	آماره تعدیل شده دایبولد-ماریانو
RMSE	۶/۸۵۷	۶/۰۳۱
MAPE	۷/۳۶۳	۶/۴۸۸

با توجه به توزیع آماری s، فرضیه قدرت پیش بینی یکسان را در سطح ۹۵٪ زمانی رد می کنیم که داشته باشیم:

$$|S| > 1/96$$

با توجه به آماره‌های بالا نتیجه می‌گیریم که عملکرد مدل PCA-SVR-PSO به طور معنی-داری بهتر از مدل SVR-PSO در پیش‌بینی شاخص کل می‌باشد.

در آزمون فرضیه دوم، عملکرد دو مدل PCA-SVR-PSO و SVR-PSO در پیش بینی شاخص ۵۰ شرکت برتر مورد آزمون قرار می‌گیرد. این فرضیه را از طریق هر دو معیار ارزیابی عملکرد میانگین قدرمطق درصد خطا و ریشه میانگین قدرمطلق خطا آزمون می‌کنیم.

$$\begin{cases} H_0 = \text{بین دو جامعه میانگین قدرمطلق درصد خطای مدلهای } SVR-PSO \text{ و } PCA-SVR-PSO \text{ اختلاف معناداری وجود ندارد} \\ H_1 = \text{بین دو جامعه میانگین قدرمطلق درصد خطای مدلهای } SVR-PSO \text{ و } PCA-SVR-PSO \text{ اختلاف معناداری وجود دارد} \end{cases}$$

جدول ٦. آزمون مقایسه زوجی قدر مطلق درصد خطا- شاخص ۵۰ شرکت

آزمون مقایسه‌های زوجی قدرمطلق درصد خطای مدل های svr-pso و pca-svr-pso		
	SVR-PSO	PCA-SVR-PSO
میانگین	۰/۸۹۲	۰/۶۸۷
واریانس	۰/۰۳۳	۰/۰۲۶
تعداد مشاهده ها	۵	
درجه آزادی	۴	
آماره t	سطح خطای ۵درصد	۵/۸۹۰
مقدار بحرانی آزمون یک طرفه		۲/۱۳۲
مقدار بحرانی آزمون دوطرفه		۲/۷۷۶
prob آزمون یک طرفه		۰/۰۰۲
آزمون دوطرفه prob		۰/۰۰۴۱۵۵۱۳۲

همان‌طور که آزمون مقایسه‌های زوجی بالا نشان می‌دهد، میانگین قدرمطلق درصد خطای مدل PCA-SVR-PSO از مدل SVR − PSO کم‌تر می‌باشد. با توجه به جدول فوق، آماره t محاسبه شده از مقدار بحرانی سطح خطا ۵ درصد بالاتر است. بنابراین فرضیه H_0 رد و فرضیه H_1 حاکی از تفاوت معنی‌دار قدرمطلق درصد خطای دو مدل به اثبات می‌رسد. در صورتی که آزمون فوق را به صورت یک طرفه انجام بدهیم، با توجه به جدول ۶، فرضیه H_0 مبنی بر بزرگ‌تر و یا مساوی بودن قدر مطلق درصد خطای مدل PCA-SVR-PSO نسبت به مدل SVR-PSO رد و فرضیه اصلی پژوهش مبنی بر بالا بودن دقت مدل PCA-SVR-PSO نسبت به مدل SVR-PSO در پیش‌بینی شاخص ۵۰ شرکت فعال‌تر تایید می‌شود.

H_0 = اختلاف معناداری وجود ندارد PCA − SVR − PSO و SVR − PSO بین دو جامعه ریشه میانگین مربع خطای مدلهای

H_1 = اختلاف معناداری وجود دارد PCA − SVR − PSO و SVR − PSO بین دو جامعه ریشه میانگین مربع خطای مدلهای

جدول ۷. آزمون مقایسه زوجی ریشه میانگین مربع خطا-شاخص ۵۰ شرکت

آزمون مقایسه‌های زوجی ریشه میانگین مربع خطای مدل های svr-pso و pca-svr-pso		
	SVR-PSO	PCA-SVR-PSO
میانگین	۱۹/۶۹۰	۱۰/۳۶۰
واریانس	۴۸/۴۵۵	۳۳/۸۵۶
تعداد مشاهده ها	۵	
درجه آزادی	۴	
آماره t	سطح خطای ۵درصد	۴/۸۰۳
مقدار بحرانی آزمون یک طرفه		۲/۱۳۲
مقدار بحرانی آزمون دوطرفه		۲/۷۷۶
prob آزمون یک طرفه		۰/۰۰۴
آزمون دوطرفه prob		۰/۰۰۸۶۳۲۰۳۶□

جدول فوق، نتایج آزمون مقایسه‌های زوجی ریشه میانگین مربع خطای مدل‌های PCA-SVR-PSO و SVR-PSO را نشان می‌دهد. نتایج جدول ۷ حاکی از پایین‌تر بودن ریشه میانگین مربع خطای مدل PCA-SVR-PSO نسبت به مدل SVR-PSO می‌باشد. جدول فوق، رد فرضیه H_0 در سطح خطای پنج

درصد را تایید می‌کند. بنابراین در سطح خطای پنج درصد، مدل PCA-SVR-PSO نسبت به مدل SVR-PSO با توجه به معیار ارزیابی عملکرد ریشه میانگین مربع خطا عملکرد بهتری در پیش بینی شاخص ۵۰ شرکت فعال‌تر دارد.

آزمون دایبولد-ماریانو برای معیارهای ارزیابی عملکرد محاسبه شده در شاخص ۵۰ شرکت فعال‌تر به صورت جدول ۸ می‌باشد.

جدول ۸. آزمون دایبولد-ماریانو شاخص ۵۰ شرکت

معیار ارزیابی عملکرد	آماره دایبولد-ماریانو	آماره تعدیل شده دایبولد-ماریانو
RMSE	۵/۳۶۹	۴/۸۰۳
MAPE	۶/۵۸۵	۵/۸۹۰

با توجه به توزیع آماری s ، فرضیه قدرت پیش بینی یکسان را در سطح ۹۵٪ زمانی رد می کنیم که داشته باشیم:

$$|S| > 1/96$$

با توجه به آماره‌های بالا نتیجه می‌گیریم که عملکرد مدل PCA-SVR-PSO به طور معنی‌داری بهتر از مدل SVR-PSO در پیش‌بینی شاخص ۵۰ شرکت فعال‌تر می‌باشد.

نتیجه‌گیری و بحث

در پژوهش حاضر فرض بر این است که پالایش اولیه داده‌ها به کاهش خطای پیش‌بینی می‌انجامد. در ابتدا مدل SVR برای پیش‌بینی شاخص کل بورس اوراق بهادار تهران در نرم‌افزار متلب شبیه سازی شده و سپس با استفاده از PCA متغیرهای ورودی به مولفه‌های اصلی تجزیه شده و به عنوان ورودی برای مدل SVR انتخاب شده است. در مدل سازی SVR، پارامترهای c و و توسط کاربر تعریف می‌شوند. پارامتر c یک پارامتر تنظیمی است و می‌تواند مقادیر صفر تا بی‌نهایت را بپذیرد. نقش این پارامتر، ایجاد تعادل میان کمینه کردن ریسک تجربی و بیشینه کردن قابلیت تعمیم‌یابی است. پارامتر نیز می‌تواند مقادیر صفر تا بی‌نهایت را بپذیرد. مقدار این پارامتر

در وضعیت بردارهای پشتیبان و در نتیجه کارایی مدل بسیار موثر است. نیز عرض کرنل با پایه شعاعی می‌باشد و مقدار آن توسط کاربر تعیین می‌شود. فرض بر این است که کاربر ممکن است در انتخاب دقیق پارامترهای مذکور دچار اشتباه شود که برای جلوگیری از این اشتباه، انتخاب پارامترهای مدل با الگوریتم بهینه سازی PSO، که یک الگوریتم قوی و جدید در حوزه بهینه‌سازی است، به نحوی انجام شده تا خطای پیش‌بینی مدل کاهش یابد. پس از برازش مدل‌ها، مقدار شاخص برای یک روز آینده تا ۲۰ روز پیش‌بینی و سپس دقت مدل‌ها در پیش‌بینی شاخص با معیارهای ارزیابی عملکرد MAPE و RMSE اندازه گیری شده است. به منظور مقایسه عملکرد مدل‌های PCA-SVR-PSO و SVR-PSO از آزمون مقایسه‌های زوجی استفاده شده که در نهایت نتایج حاکی از بالاتر بودن دقت پیش‌بینی مدل PCA-SVR-PSO نسبت به مدل SVR-PSO بود.

بنابراین نتیجه می‌گیریم که انتخاب صحیح ورودی‌ها و کاهش بعد داده‌ها می‌تواند باعث بهبود عملکرد ماشین‌های بردار پشتیبان در پیش‌بینی سری زمانی شاخص بورس اوراق بهادار تهران شود.

منابع

- بروکز کریس، ۱۳۸۹، اقتصاد سنجی مالی و تجزیه و تحلیل داده‌ها در علوم انسانی، ترجمه بدری احمد و عبدالباقی عبدالمجید، جلد۱، موسسه عملی فرهنگی نص، چاپ اول.
- جانسون، ریچارد. ۱۳۷۸، تحلیل آماری چند متغیری کاربردی، ترجمه نیرومند، حسینعلی. دانشگاه فردوسی مشهد، چاپ سوم.
- کیانی، محمد فریدون. میرعرب شاهی، رامین. حسین خانی، ابراهیم. ۱۳۸۷،تعیین میزان دقت دسته‌بندی کننده ماشین بردارپشتیبان در ارزیابی اعتباری بانکی، دومین همایش ملی مهندسی برق کامپیوتر و فناوری اطلاعات، صص ۲۰۰-۲۰۸.
- نوری، روح‌اله. خاکپور، امیر. دهقانی، مجید. فرخ‌نیا، اشکان. ۱۳۸۹، پیش‌بینی ماهانه جریان آب با استفاده از ماشین بردارپشتیبان بر مبنای آنالیز مولفه اصلی، فصلنامه علمی و پژوهشی آب و فاضلاب، دوره۲۲، شماره۱، صص ۱۱۸-۱۲۳
- Chi-Jie, Lu. (2013). Hybridizing nonlinear independent component analysis and support vector regression with particle swarm optimization for stock index forecasting, Neural Applied & Soft computing, volume 40, pp. 164-178.
- Gunn, steve .(1998). Support Vector Machines for Classification and Regression, university of SOUTHAMPTON, chapter 2, pp. 5-16.
- Kenndy, J. Eberhart, R.C.(1995). Particle Swarm Optimization. In Proceedings of the IEE International Conference on Neural Networks IV.
- Li. Xia, He. Kaijjian.(2014). Forecasting Crude Oil Price With Multiscale Denoising Ensemble Model, Mathematic Problems in Engineering, pp.1-19
- Lipo, W. (2005), Support Vector Machines, Theory and Application, university of Auckland.
- Pei. Chann Chang, Jhen. Wu, Jyun. Lin.(2015), A Takagi-Sugeno fuzzy model combined with a support vector regression for stock trading forecasting, Applied Soft Computing, volume 38, pp. 831-842.
- Sanches, D. (2003). Advanced Support Vector Machines and Kernel methods, Neurocomputing, volume 55, pp. 5-20.