

The Index Prediction of Tehran Stock Exchange by Combining the Principal Components Analysis, Support Vector Regression and Particle Swarm Optimization

Reza Raei¹
Ali Nikahd Ghoseiraei²
Mostafa Habibi³

Abstract

The prediction of future fluctuations in stock index can provide information about the future trend in the capital market. In order to increase the accuracy of the prediction of stock exchange index, this study used a combination of statistical methods and artificial intelligence. In this study, a hybrid stock index prediction model by utilizing principal component analysis (PCA), support vector regression (SVR) and particle swarm optimization (PSO) is proposed. In the proposed model, first, the PCA is used to deal with the nonlinearity property of the stock index data. The proposed model utilizes PCA to extract features from the observed stock index data. The features which can be used to represent underlying/hidden information of the data are then served as the inputs of SVR to build the stock index prediction model. Finally, PSO is applied to optimize the parameters of the SVR prediction model since the parameters of SVR must be carefully selected in establishing an effective and efficient SVR model. The findings show that preprocessing the data can decrease the prediction error of the model significantly.

Keywords: Stock Index, Principal Component Analysis, Support Vector Regression, Particle Swarm Optimization, Prediction.

JEL: G10, G19, G17, C02

-
- 1 . Department of Financial Management, Faculty of Management, University of Tehran, Tehran, Iran, Email: raei@ut.ac.ir
 - 2 . Department of Financial Management, Faculty of Management, University of Tehran, Tehran, Iran, Email: h.nikahd@gmail.com
 - 3 . Department of Financial Management, Faculty of Management, University of Tehran, Tehran, Iran, Corresponding Author, Email: mostafahabibi_68@yahoo.com

<http://jfm.alzahra.ac.ir/>

پیش بینی شاخص بورس اوراق بهادار تهران با ترکیب روش های آنالیز

مولفه های اصلی، رگرسیون بردار پشتیبان و حرکت تجمعی ذرات^۱

رضا راعی^۲، علی نیک عهد قصرائی^۳ و مصطفی حبیبی^۴

چکیده

پیش بینی نوسان های آینده شاخص سهام می تواند اطلاعاتی در مورد روند آینده بازار سرمایه فراهم نماید. در این پژوهش، به منظور افزایش دقت پیش بینی شاخص بورس اوراق بهادار تهران، ترکیبی از روش های آماری و هوش مصنوعی به کار رفته است. مدل اصلی پیش بینی در این پژوهش، رگرسیون بردار پشتیبان بهینه شده به وسیله الگوریتم حرکت تجمعی ذرات می باشد. 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 زمان بر بودن و کارایی پایین انتخاب پارامتر توسط کاربر، برای انتخاب ترکیب بهینه پارامترهای مدل رگرسیون بردار پشتیبان، از روش بهینه سازی حرکت تجمعی ذرات استفاده شده است که الگوریتمی قوی در حوزه بهینه سازی می باشد. با توجه به حجم زیاد داده های ورودی به مدل برای کاهش زمان یادگیری و افزایش دقت پیش بینی، با استفاده از روش آنالیز مولفه های اصلی، پیش پردازش روی متغیرهای ورودی صورت گرفته و به مولفه های اصلی تبدیل شده است. نتایج بدست آمده نشان داد که پیش پردازش روی داده ها، خطای پیش بینی مدل را به طور قابل ملاحظه ای کاهش داده است.

واژه های کلیدی: شاخص بورس، آنالیز مولفه های اصلی، رگرسیون بردار پشتیبان، بهینه سازی حرکت

تجمعی ذرات، پیش بینی

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

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

مقدمه

بورس اوراق بهادار، مکانی است که در آن پس اندازهای راكد جمع آوری شده و در تامین مالی پروژه‌های سرمایه‌گذاری بلند مدت استفاده می‌شود. افراد دارای پس اندازهای راكد، بازدهی سرمایه‌گذاری در بورس را با سایر گزینه‌های سرمایه‌گذاری (سرمایه‌گذاری در بخش مسکن، سرمایه‌گذاری در بانک، طلا، تولید مستقیم و ...) مقایسه کرده و با توجه به ریسک و بازده هر یک از سرمایه‌گذاری‌های ممکن، تصمیم‌گیری 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 از آن به عنوان یک متغیر پیشرو برای پیش‌بینی نوسان فعالیت‌های اقتصادی استفاده کرد. اما از آنجایی 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) برای پالایش اولیه داده‌ها استفاده شده است. سپس با استفاده از رگرسیون بردار پشتیبان (SVR)^۲ که نوع خاصی از ماشین‌های بردار پشتیبان (SVM)^۳ می‌باشد، به پیش‌بینی شاخص اقدام شده است. در نهایت با استفاده از روش بهینه‌سازی حرکت جمعی ذرات (PSO)^۴ که یک روش بهینه‌سازی تکاملی می‌باشد، پارامترهای مدل SVR 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 (تقریباً) 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 در مورد k مولفه اصلی کاهش داد.

تحلیل مولفه‌های اصلی وسیله‌ای برای رسیدن به هدف هستند تا این که 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, \dots, 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, \dots, X_p به عنوان محورهای مختصات به دست 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 ρ) X_1, X_2, \dots, X_p مربوط می‌شود. برای گسترش آن‌ها، گسترش نرمال چند متغیری لازم نیست. از سوی دیگر، مولفه‌های اصلی جامعه‌های نرمال چند متغیری، تعبیر 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 $X' = [X_1, X_2, \dots, X_p]$ دارای ماتریس کوواریانس Σ با مقادیر ویژه $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ است.

ترکیبات خطی زیر را در نظر می‌گیریم:

$$\begin{aligned} Y_1 &= e'_1 X = e_{11}X_1 + e_{21}X_2 + \dots + e_{p1}X_p \\ Y_2 &= e'_2 X = e_{12}X_1 + e_{22}X_2 + \dots + e_{p2}X_p \\ Y_p &= e'_p X = e_{1p}X_1 + e_{2p}X_2 + \dots + e_{pp}X_p \end{aligned} \quad (1)$$

در این صورت داریم:

$$\text{Var}(Y_i) = e'_i \Sigma e_i \quad i = 1, 2, \dots, p \quad (2)$$

$$\text{Cov}(Y_i, Y_k) = e'_i \Sigma e_k \quad i, k = 1, 2, \dots, p \quad (3)$$

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

نخستین مولفه اصلی، یک ترکیب خطی با واریانس ماکزیمم است. یعنی $Var(Y_i) = e_i' \Sigma e_i$ را ماکزیمم می‌کند. واریانس دومین مولفه اصلی، کمتر از مولفه نخست می‌باشد. واریانس مولفه‌ها کاهش می‌یابد تا این که مولفه p ام کمترین واریانس را دارا دارد. از تجزیه و تحلیل مولفه‌های اصلی نتایج زیر حاصل می‌شود:

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

$$Y_i = e_i' X = e_{1i} X_1 + e_{2i} X_2 + \dots + e_{pi} X_p, \quad i = 1, 2, \dots, p \quad (۴)$$

داده شود. با این انتخاب‌ها داریم:

$$Var(Y_i) = e_i' \Sigma e_i = \lambda_i \quad i = 1, 2, \dots, p \quad (۵)$$

$$Cov(Y_i, Y_k) = e_i' \Sigma e_k = 0 \quad i \neq k \quad (۶)$$

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

نتیجه ۲- فرض کنید فرض کنید Σ ماتریس کوواریانس بردار تصادفی $X' = [X_1, X_2, \dots, X_p]$ باشد. فرض کنید Σ دارای زوج مقدار ویژه-بردار ویژه $(\lambda_1, e_1), (\lambda_2, e_2), \dots, (\lambda_p, e_p)$ باشد که $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ همچنین فرض کنید $Y_1 = e_1' X, Y_2 = e_2' X, \dots, Y_p = e_p' X$ مولفه‌های اصلی باشند که در آن صورت داریم:

$$\begin{aligned} \sigma_{11} + \sigma_{22} + \dots + \sigma_{pp} &= \sum_{i=1}^p Var(X_i) \\ &= \lambda_1 + \lambda_2 + \dots + \lambda_p \end{aligned} \quad (۷)$$

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

$$\left(\begin{array}{l} \text{سهم کل واریانس} \\ \text{جامعه مربوط} \\ \text{به مولفه اصلی} \\ \text{م } k \end{array} \right) = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \dots + \lambda_p} \quad k = 1, 2, \dots, p \quad (8)$$

نتیجه ۳- اگر $Y_1 = e_1'X, Y_2 = e_2'X, \dots, Y_p = e_p'X$ ، مولفه‌های اصلی به دست آمده از ماتریس کوواریانس Σ باشد. آن گاه:

$$\rho_{Y_i, X_k} = \frac{e_{ki} \sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \quad i, k = 1, 2, \dots, p \quad (9)$$

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

روش ماشین بردار پشتیبان، یکی از روش‌های یادگیری ماشینی است که بر مبنای تئوری یادگیری آماری وپنیک در دهه ۹۰ میلادی توسط وپنیک و همکارانش عرضه شد. در SVM، از اصول کمینه‌سازی ریسک ساختاری (SRM) استفاده شده است، در حالی که سایر روش‌ها از اصول کمینه‌سازی ریسک تجربی (ERM) 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. Empirical Risk Minimization

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

۳- ماشین بردار پشتیبان، با کمینه کردن ریسک ساختاری، بهترین جواب را می دهد (سنچس، ۲۰۰۳).

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

$$L_\varepsilon(X, Y, f(x)) = \begin{cases} |y - f(x)| - \varepsilon & |y - f(x)| \leq \varepsilon \\ 0 & \text{other wise} \end{cases} \quad (10)$$

$L_\varepsilon(y)$ معرف تابع ضرر و ε خطای مجاز در تابع ضرر می باشد. پارامترهای کنترل کننده تابع رگرسیون بهینه با حل مساله بهینه سازی زیر به دست می آیند.

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

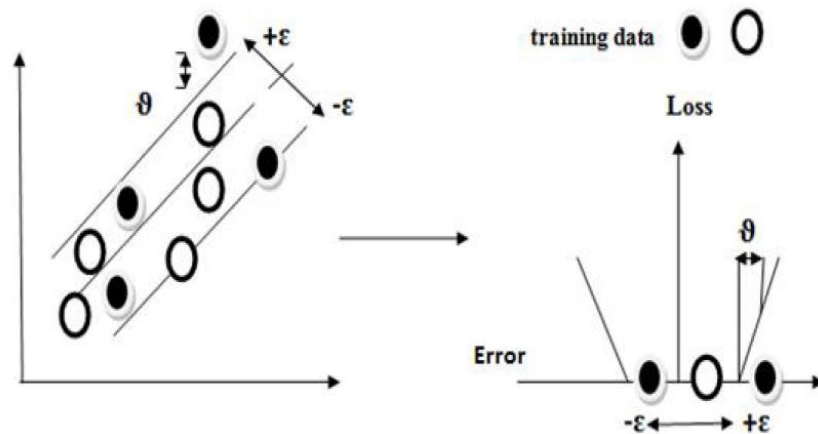
subject to:

$$\left((W \cdot X_j) + b \right) - y_j \leq \varepsilon + \vartheta_j^* \quad (11)$$

$$y_j - \left((W \cdot X_j) + b \right) \leq \varepsilon + \vartheta_j$$

$$\vartheta_j, \vartheta_j^* \geq 0$$

در رابطه ۱۱، θ و θ^* متغیرهای slack هستند. این متغیرها به همراه تابع ضرر در شکل ۱ نشان داده شده‌اند (گان، ۱۹۹۸).



شکل ۱. تابع ضرر وینیک و متغیرهای slack

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

$$L(a^*, a) = -\varepsilon \sum_{i=1}^k (a_i^* + a_i) + \sum_{i=1}^k y_i (a_i^* - a_i) - \frac{1}{2} \sum_{i=1}^k \sum_{j=1}^k (a_i^*, a_i) (a_j^*, a_j) (x_i, x_j) \quad (12)$$

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

$$\begin{cases} \sum a_i^* = \sum a_i \\ 0 \leq a_i^* \leq C \\ 0 \leq a_i \leq C \end{cases} \quad \text{for } i = 1, 2, 3, \dots, k \quad (13)$$

مساله بهینه سازی فوق به کمک روش های برنامه ریزی درجه دو (QP) قابل حل می باشد. در نتیجه رسیدن به اکسترمم کلی نیز قطعی خواهد بود و خطر به دام افتادن در اکسترمم محلی وجود ندارد. داده هایی که ضرایب لاگرانژ متناظر با آنها غیر صفر باشد، به عنوان بردار پشتیبان شناخته می شوند. این داده ها از نظر هندسی دارای خطای پیش بینی بزرگتر از $\bar{\epsilon}$ هستند. بنابراین بردارهای پشتیبان درون باند $\bar{\epsilon}$ قرار نمی گیرند و مقدار ϵ ، تعداد بردارهای پشتیبان را کنترل می کند.

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

$$W_0 = \sum (a_i^* - a_i)x_i \quad (14)$$

$$b_0 = -\left(\frac{1}{2}\right) \cdot W_0 \cdot [x_r + x_s] \quad (15)$$

$$f(x) = \sum (a_i^* - a_i)(x_i - x) + b_0 \quad (16)$$

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

برای ساخت مدل ماشین بردار پشتیبان، پارامترهای c و ϵ توسط کاربر تعریف می شوند. پارامتر c یک پارامتر تنظیمی است و می تواند مقادیر صفر تا بی نهایت را بپذیرد. نقش این پارامتر، ایجاد تعادل میان کمینه کردن ریسک تجربی و بیشینه کردن قابلیت تعمیم یابی است. پارامتر ϵ نیز می تواند 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 کنترل‌کننده تابع بهینه با روابط زیر محاسبه می‌شوند (سنجس، ۲۰۰۳)

$$W \cdot X = \sum (a_i^* - a_i) K(x_i, x) \quad (17)$$

$$b_0 = -\frac{1}{2} \sum (a_i^* - a_i) [K(x_r, x_i) + K(x_s, x_i)] \quad (19)$$

در این روابط $K(x_i, x)$ نشانگر تابع کرنل می‌باشد. موجودات طبیعی گاهی به صورت یک توده رفتار می‌کنند. یکی از جریانهای اصلی در پژوهش زندگی مصنوعی، بررسی چگونگی رفتار موجودات طبیعی به صورت یک توده و پیاده‌سازی دوباره مدل توده‌ها در رایانه است.

یک روش جدید بهینه‌سازی با استفاده از همانند سازی رفتار گروهی موجودات طبیعی در اوایل دهه ۱۹۹۰ ابداع شد. ابرهارت و کندی (۱۹۹۵)، بهینه‌سازی ذرات انبوه (PSO) را بر اساس شبیه‌سازی از توده‌های پرندگان و دسته‌ماهی‌ها توسعه دادند. هر فرد، تجارب قبلی خود را در PSO مبادله می‌کند. PSO برای حل مسایل بهینه‌سازی غیرخطی با متغیرهای پیوسته ایجاد شده است. علاوه بر این، بر خلاف روش‌های تکاملی دیگر مانند الگوریتم ژنتیک، PSO می‌تواند با تنها یک برنامه کوچک پیاده‌سازی شود. این قابلیت PSO، یکی از مزیت‌های آن در مقایسه با دیگر تکنیک‌های بهینه‌سازی است. PSO روشی مبتنی بر تکنیک‌های تصادفی است که از آن می‌توان برای پیدا کردن مینیمم سراسری (غیرقطعی) مسایل برنامه‌ریزی غیرخطی استفاده کرد.

کندی و ابرهارت، PSO را از طریق شبیه‌سازی دسته پرندگان توسعه دادند. موقعیت هر عامل با s و همچنین سرعت آن با v نمایش داده می‌شود. اصلاح موقعیت عامل با استفاده از اطلاعات موقعیت و سرعت صورت می‌گیرد.

دسته پرندگان، یک تابع هدف خاص را بهینه‌سازی می‌کنند. هر عامل بهترین مقدار تجربه کرده را $pbest$ و موقعیت فعلی را s می‌داند. این اطلاعات، تجربیات شخصی هر عامل است. علاوه بر

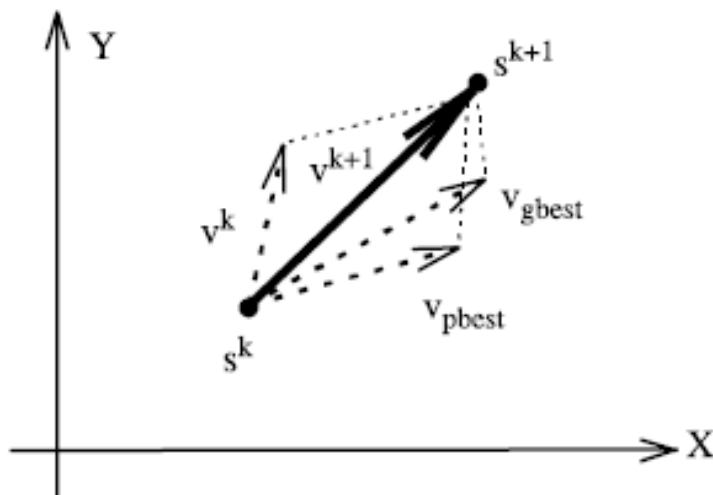
این، هر عامل، بهترین مقدار به دست آمده در گروه را $gbest$ می داند. هر عامل تلاش می کند تا موقعیت خود را با استفاده از اطلاعات زیر تغییر دهد:

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:

$$v_i^{k+1} = wv_i^k + c_1randd_1 \times (pbest_i - s_i^k) + c_2rand_2 \times (gbest - s_i^k) \quad (20)$$

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.

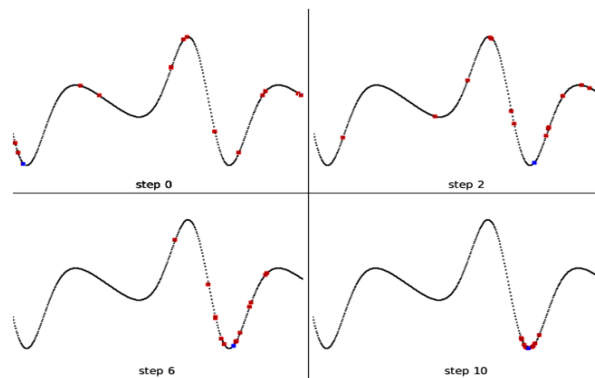


شکل ۲

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

$$S_i^{k+1} = S_i^k + v_i^{k+1} \quad (21)$$

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 می‌کند و نقاط جستجو، تدریجی به نقطه بهینه با استفاده از pbest و gbest نزدیک می‌شوند.



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 به دست 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 به دست آمده می‌توان به این نتیجه رسید که ماشین بردار پشتیبان، یک روش جدید و قابل اطمینان در میان دیگر روش‌های داده کاوی است.

چی جی‌لو و همکاران (۲۰۱۳) در مقاله‌ای تحت عنوان "پیش‌بینی شاخص با مدل هیبریدی آنالیز مولفه‌های مستقل و رگرسیون بردار پشتیبان 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 مویک 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 بردار پشتیبان 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 مدل‌های 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.

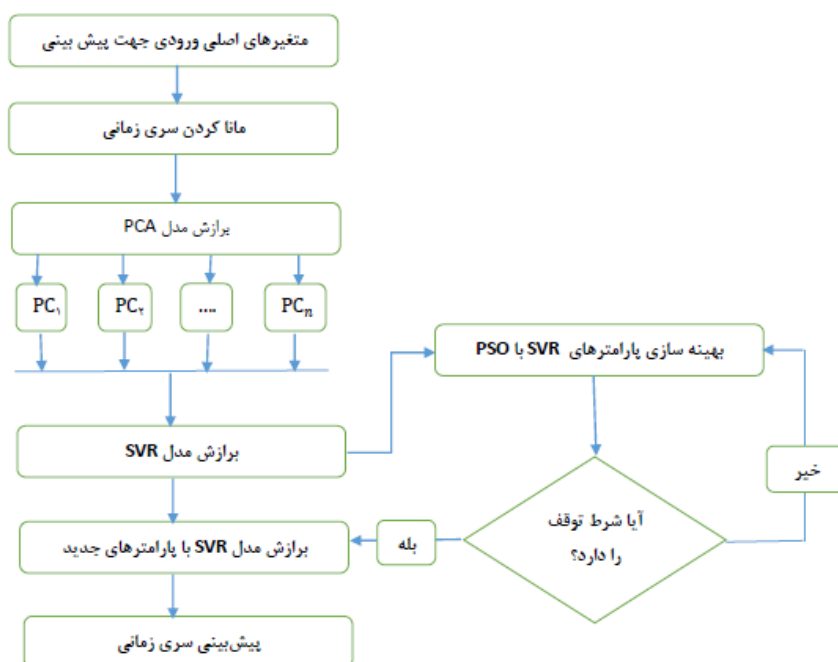
به طور خلاصه سازوکار اصلی SVM در حل مساله رگرسیون به صورت زیر بیان می‌شود:

- 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 ϵ مجاز است (تابع ضرر).

- ۳- ماشین بردار پشتیبان، با کمینه کردن 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 ϵ و σ 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 ϵ نیز می‌تواند مقادیر صفر تا بی‌نهایت را بپذیرد. مقدار این پارامتر در وضعیت بردارهای پشتیبان و در نتیجه کارایی مدل بسیار موثر است. σ 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 ۴. 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 غلطان به پنج دوره زمانی برابر تقسیم شده و سپس برای هر دوره پیش‌بینی با مدل‌های SVR-PSO (استفاده از داده‌های معمولی برای پیش‌بینی 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

با برآزش 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

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

در آزمون فرضیه نخست، عملکرد دو مدل PCA-SVR-PSO و SVR-PSO در پیش‌بینی شاخص کل مورد آزمایش قرار می‌گیرد. این فرضیه را از طریق هر دو معیار ارزیابی عملکرد میانگین قدرمطلق درصد خطا و ریشه میانگین قدرمطلق خطا آزمون می‌کنیم.

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

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

آزمون مقایسه‌های زوجی قدرمطلق درصد خطای مدل های svr-psy و pca-svr-psy		
PCA-SVR-PSO	SVR-PSO	
۰/۴۸۶	۰/۷۵۳	میانگین
۰/۰۸۹	۰/۰۶۹	واریانس
۵		تعداد مشاهده‌ها
۴		درجه آزادی
۶/۴۸۸	سطح خطای ۵ درصد	آماره t
۲/۱۳۲		مقدار بحرانی آزمون یک طرفه
۲/۷۷۶		مقدار بحرانی آزمون دو طرفه
۰/۰۰۱		prob آزمون یک طرفه
۰/۰۰۲۹۰۹۹۵۶۰		آزمون دو طرفه prob

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.

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

جدول ۴. آزمون مقایسه زوجی ریشه میانگین مربع خطا-شاخص کل

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

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 پایین تر 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.

همان گونه که گفته شد، علاوه بر آزمون مقایسه‌های زوجی برای بررسی معناداری تفاوت بین میانگین دو جامعه ریشه میانگین مجذور خطا و درصد قدر مطلق خطا، از دو آماره دایبولد-ماریانو و آماره هاروی-لیورن-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 پیش‌بینی یکسان را در سطح 95% زمانی رد می‌کنیم 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 کل می‌باشد.

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

بین دو جامعه میانگین قدرمطلق درصد خطای مدل‌های SVR-PSO و PCA-SVR-PSO اختلاف معناداری وجود ندارد = H_0
 بین دو جامعه میانگین قدرمطلق درصد خطای مدل‌های SVR-PSO و PCA-SVR-PSO اختلاف معناداری وجود دارد = H_1

جدول ۶. آزمون مقایسه زوجی قدرمطلق درصد خطا- شاخص ۵۰ شرکت

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

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 قدرمطلق 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.

بین دو جامعه ریشه میانگین مربع خطای مدل‌های SVR – PSO و PCA – SVR – PSO اختلاف معناداری وجود ندارد $H_0 =$
 بین دو جامعه ریشه میانگین مربع خطای مدل‌های SVR – PSO و PCA – SVR – PSO اختلاف معناداری وجود دارد $H_1 =$

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

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

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 پیش‌بینی 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

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

با توجه به توزیع آماری S، فرضیه قدرت پیش‌بینی یکسان را در سطح ۹۵٪ زمانی رد می‌کنیم 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 شبیه‌سازی 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 E و σ 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 E نیز می‌تواند مقادیر صفر تا بی‌نهایت را بپذیرد. مقدار این پارامتر

در وضعیت بردارهای پشتیبان و در نتیجه کارایی مدل بسیار موثر است. σ 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 بهینه‌سازی 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 اندازه‌گیری 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، جلد ۱، 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 بردار پشتیبان 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 بردار پشتیبان بر مبنای آنالیز مولفه اصلی، فصلنامه علمی و پژوهشی آب و فاضلاب، دوره ۲۲، شماره ۱، 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. Kaijian. (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.