مدل رگرسیون بردار تکیه‌گاه و مقایسه آن با رگرسیون نیم‌پارامتری

مهدی روزبه[۱]، آرتا روحی[۲]، فاطمه جهادی [۳] و سعید زال‌زاده[۴]

تاریخ دریافت: ۱۴۰۰/۰۴/۱۳
تاریخ پذیرش: ۱۴۰۰/۱۲/۲۶

چکیده:

در این تحقیق، هدف بررسی و تحلیل روشی برای پیش‌بینی قیمت سهام بورس اوراق بهادار است. هرچند پیش‌بینی بازار سرمایه با توجه به وابستگی آن به عامل سیاست چندان ساده نیست، اما با مدل‌سازی داده‌ها، پیش‌بینی عملکرد سهام بورس اوراق بهادار در بازه بلندمدت تا حدودی امکان‌پذیر خواهد بود. در این راستا با استفاده از مدل‌های رگرسیون نیم‌پارامتری و رگرسیون بردار تکیه‌گاه با هسته‌های مختلف و اندازه‌گیری خطاهای پیش‌بین، بر روی یکی از سهم‌های بازار بورس اوراق بهادار بر اساس نوسان‌های روزانه و مقایسه روش‌ها با استفاده از معیارهای ریشه میانگین توان دوم خطاها و میانگین قدرمطلق درصد خطاها، مدل رگرسیون بردار تکیه‌گاه با هسته شعاعی و خطای برابر ۰/۱ دارای مناسب‌ترین برازش روی داده‌های واقعی بازار سهام بوده است.

واژه‌های کلیدی: پیش‌بینی سهام، مدل رگرسیون بردار تکیه‌گاه، مدل رگرسیون نیم‌پارامتری، یادگیری ماشین.

۱ مقدمه

پیشرفت محاسباتی منجر به ایجاد چندین الگوریتم یادگیری ماشین شده است که برای پیش‌بینی حرکت مداوم بازار به کار می‌رود و بنابراین می‌توان ارزش دارایی‌های آینده مانند قیمت سهام را برآورد کرد [۹]. مدل‌های مبتنی بر ماشین بردار تکیه‌گاه [۵] از جمله تکنیک‌های بسیار پرکاربرد در این زمینه هستند. سیستم‌های معاملات با توجه به ویژگی‌های سری زمانی مالی، هنگامی‌که گسترش می‌یابند با چالش‌های مختلفی روبرو می‌شوند [۸].

یادگیری ماشین شاخه جدیدی از تحلیل آماری است که از قدرت محاسباتی گسترده‌ای توسط رایانه‌ها برای تجزیه داده‌های بزرگ برخوردار است، استفاده می‌کند. یکی از زمینه‌هایی که بسیار موردتوجه قرار می‌گیرد پیش‌بینی بازار سهام با استفاده از الگوریتم‌های یادگیری ماشین است. طبقه‌بندی‌ها، سیستم‌هایی هستند که می‌توانند از طریق آموزش یاد بگیرند و الگوهایی را بشناسند و داده‌ها را به کلاس جدید اختصاص دهند. [۱۰، ۱۹].

برخی از محققان از الگوریتم‌های یادگیری ماشین برای افزایش عملکرد پیش‌بینی استفاده کرده‌اند. پیشنهاد می‌شود شبکه‌های عصبی مختلف با ماشین‌های یادگیری تکیه‌گاه ادغام شوند تا بتوان ارزش روزانه سهام را پیش‌بینی کرد.

روش ماشین بردار تکیه‌گاه یکی از روش‌های برآورد در یادگیری ماشین است. در این روش ابتدا داده‌ها را به دو قسمت تقسیم می‌شود که قسمت اول مربوط به داده‌های آموزش است. ۷۰ درصد کل داده‌ها را به‌صورت تصادفی انتخاب می‌کنیم و مدلی که بهترین برازش را روی داده‌ها دارد انتخاب می‌شود و در ادامه مدل را روی ۳۰ درصد باقی داده‌ها امتحان می‌کنیم اگر نتیجه رضایت‌بخش بود مدل انتخاب‌شده تائید می‌گردد، در غیر این صورت مجدد به دنبال یک مدل جدید خواهیم رفت.

رگرسیون نیم‌پارامتری ، کاربرد زیادی در مسائل اقتصادی دارد که از مهم‌ترین آن‌ها می‌توان به مدل تابع درآمد سرمایه انسانی [۱۸] و

[۱] هیئت‌علمی گروه آمار، دانشگاه سمنان، سمنان، ایران (نویسنده مسئول mahdi.roozbeh@semnan.ac.ir).
[۲] دانش‌آموخته کارشناسی ارشد آمار، دانشگاه سمنان، سمنان، ایران.
[۳] دانش‌آموخته کارشناسی ارشد آمار، دانشگاه سمنان، سمنان، ایران.
[۴] هیئت‌علمی گروه آمار، دانشگاه سمنان، سمنان، ایران.
[۴] کد موضوع‌بندی ریاضی (۲۰۱۰): 62J05.

[5] Support Vector Machines

منحنی دستمزد [۴] اشاره کرد که مدل فوق ترکیبی از مدل خطی و مدل ناپارامتری می‌باشد. مدل‌های رگرسیون نیم‌پارامتری نخستین بار توسط انگل و همکاران [۷] در سال ۱۹۸۶ در بررسی رابطه میان مصرف ماهیانه برق به‌عنوان متغیر پاسخ و قیمت ماهیانه برق، درآمد ماهیانه و دمای هوا به‌عنوان متغیرهای پیشگو با در نظر گرفتن قیمت ماهیانه برق و درآمد ماهیانه به‌عنوان بخش خطی و دمای هوا به‌عنوان بخش غیرخطی مورداستفاده قرار گرفتند. برای به دست آوردن پارامترها و شیوه‌های برآورد در این روش می‌توان به روزبه و امینی [۱] مراجعه کرد.

بر اساس مطالعات نایاک و همکاران [۱۲]، پتال و همکاران [۱۳]، راجو و همکاران [۳] در سال ۲۰۱۵، ماناهوف و همکاران [۱۱] و چودهاری و همکاران [۵] در سال ۲۰۱۴ معیارهایی که برای ارزیابی مدل موردبررسی قرار می‌گیرند عبارت‌اند از ریشه میانگین توان‌های دوم خطاها [۶] (فرمول سمت چپ) و استفاده از میانگین قدرمطلق درصد خطاها [۷] (فرمول سمت راست)، که فرمول‌های محاسباتی آن‌ها به‌صورت زیر است:

$$R=\sqrt{\frac{۱}{T}\sum_{i=۱}^{T}(d_i-\hat{d}_i)^۲},\quad M=\frac{۱}{T}\sum_{i=۱}^{T}\left|\frac{d_i-\hat{d}_i}{d_i}\right|$$

به‌طوری‌که در فرمول بالا d_i مقدار واقعی داده‌ها، $\hat{d}_i$ مقدار برازش شده و T نمونه‌های آزمون کل می‌باشد. بدیهی است که روش فوق با معیار انتخابی کمتر مناسب‌تر است.

۲ ماشین بردار تکیه‌گاه

۱.۲ مسئله بهینه‌سازی حاشیه سخت در یادگیری ماشین‌های بردار تکیه‌گاه

روش ماشین بردار تکیه‌گاه توسط وپنیک [۱۷] در سال ۱۹۹۵ به وجود آمده است، به‌طورمعمول برای طبقه‌بندی مشاهدات بین گروه‌ها استفاده می‌شود. نقطه‌ای مانند x_n به‌گونه‌ای در نظر گرفته می‌شود که نزدیک‌ترین نقطه به ابرصفحه باشد و آن را بر روی خط می‌توان قرار داد. با در نظر گرفتن ابرصفحه $|W^\top x_n+w_\circ|=\circ$ برای $W=[w_۱,\ldots,w_d]$، معادلات خطوط حاشیه‌ای به شکل $|W^\top x_n+w_\circ|=۱$ می‌باشد. هدف این است که حاشیه تا حدی زیاد شود که فاصله‌ی خط حاشیه تا نقطه x_n به کمترین مقدار برسد. اکنون با داشتن فرمول فاصله و جایگذاری مقادیر به‌دست‌آمده در آن، فاصله بالا به‌صورت زیر تعریف می‌گردد:

$$D=\frac{|W^\top x_n+w_\circ|}{\|W\|}=\frac{۱}{\|W\|},\qquad (۱)$$

به دلیل وجود دو حاشیه در اطراف خط موردنظر، بنابراین هدف بیشینه کردن $\frac{۲}{\|W\|}$ می‌باشد، حل این مسئله (با توجه به‌شرط آن $\min_{W,w_\circ}|W^\top x_n+w_\circ|=۱$) دشوار است. لذا با توجه به مقادیری که $y_n=+۱$ و $y_n=-۱$ می‌گیرند، می‌توان نوشت:

$$\min_{W,w_\circ}\frac{۱}{۲}\|W\|^۲,\qquad s.t.\quad y_n(W^\top x_n+w_\circ)\geq ۱$$

مسئله فوق یک مسئله برنامه‌نویسی درجه دوم [۸] (QP) که معادله آن به‌صورت زیر است:

$$\min_x \frac{۱}{۲}x^\top Qx^\top+c^\top x,\qquad s.t.\quad Ax\leq b,\quad Ex=b,\quad (۲)$$

حل مسئله فوق به‌صورت مسئله برنامه‌نویسی درجه دوم دشوار است درنتیجه ابتدا با تعریف تابع لاگرانژ چند متغیره برای بهینه‌سازی داریم:

$$p^*=\min_x f(x),\qquad s.t.\begin{cases}g_i(x)\leq \circ, & i=۱,\ldots,m\\ h_i(x)=\circ, & i=۱,\ldots,p\end{cases}$$

$$\ell(x,\alpha,\lambda)=f(x)+\sum\alpha_i g_i(x)+\sum\lambda_i h_i(x),\qquad (۳)$$

با در نظر گرفتن $\alpha_i\geq\circ$ و مقادیر λ_i حالت‌های مختلفی از تابع لاگرانژ فوق به دست می‌آید:

$$\max_{\alpha_i\geq\circ,\lambda_i}\ell(x,\alpha,\lambda)=\begin{cases}\infty, & \forall g_i(x)>\circ\\ \infty, & \forall h_i(x)\neq\circ\\ f(x), & \text{اگر شرایط برقرار باشد}\end{cases},$$

حال با توجه به فرمول‌های فوق و مسئله دوگان $p^*=\min_x\max_{\alpha_i\geq\circ,\lambda}\ell(x,\alpha,\lambda)$ است. با توجه به موارد گفته‌شده می‌توان نوشت:

$$\max_x\min_y h(x,y)\leq\min_y\max_x h(x,y),$$

$$p^*=\min_x\max_{\alpha_i\geq\circ,\lambda_i}\ell(x,\alpha,\lambda),\quad d^*=\max_{\alpha_i\geq\circ,\lambda_i}\min_x\ell(x,\alpha,\lambda),$$

$$d^*\leq p^*\qquad (۴)$$

[6] Root Mean Squared Error
[7] Mean Absolute Percentage Error
[8] Quardratic Programming

اگر به‌جای p^*، عبارت $\min_{W,w_\circ} \frac{1}{2}\|W\|^2$ قرار گیرد، داریم:

$$\min_{W,w_\circ} \max_{\alpha_n \geq \circ} \{\frac{1}{2}\|W\|^2 + \sum \alpha_n (1 - y_n(W^\top x_n + w_\circ))\}, \quad (۵)$$

$$\max_{\alpha_n \geq \circ} \min_{W,w_\circ} \{\frac{1}{2}\|W\|^2 + \sum \alpha_n (1 - y_n(W^\top x_n + w_\circ))\}, \quad (۶)$$

با توجه به معادلات فوق و مسئله دوگان معادله (۵) بزرگ‌تر یا مساوی معادله (۶) است، برای کمینه کردن قسمت داخلی معادله (۶) باید از آن مشتق گرفت و برابر صفر قرار داد:

$$\max_{\alpha_n \geq \circ} \min_{W,w_\circ} \underbrace{\{\frac{1}{2}\|W\|^2 + \sum \alpha_n (1 - y_n(W^\top x_n + w_\circ))\}}_{\text{با مشتق‌گیری و برابر صفر قرار دادن}}, \quad (۷)$$

با مشتق گرفتن نسبت به W و $w_\circ$ می‌توان نوشت:

$$\nabla_W \ell(W, w_\circ, \alpha) = W - \sum \alpha_n y_n x_n = \circ \Rightarrow W = \sum \alpha_n y_n x_n$$
$$\frac{\partial \ell(W, w_\circ, \alpha)}{\partial w_\circ} = \circ \Rightarrow -\sum \alpha_n y_n = \circ$$

با قرار دادن مقادیر به‌دست‌آمده در رابطه (۷) خواهیم داشت:

$$\min_{W,w_\circ} \ell(W, w_\circ, \alpha) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m x_n x_m,$$
$$\sum_{n=1}^{N} \alpha_n y_n = \circ, \alpha_n \geq \circ$$

اکنون باید تابع بالا را بیشینه کرد. چون α درجه دوم می‌باشد دوباره مسئله برنامه‌نویسی درجه دوم مطرح است، یعنی داریم:

$$x = \alpha_n, \quad Q = \begin{bmatrix} y_1 y_1 x_1^\top x_1 & \dots & y_1 y_N x_1^\top x_N \\ \vdots & \ddots & \vdots \\ y_N y_1 x_N^\top x_1 & \dots & y_N y_N x_n^\top x_N \end{bmatrix}, \quad c^\top = 1^\top$$

با توجه به شروط $y^\top \alpha = \circ, \alpha > \circ$ و با جایگذاری مقادیر فوق در معادله مسئله برنامه‌نویسی درجه دوم (۲) و یافتن α هایی که تابع (۸) را بیشینه می‌کنند. با جایگذاری در $W = \sum \alpha_n y_n x_n$، مقدار W را می‌توان به دست آورد، برای یافتن مقدار $w_\circ$ از شرایط کاروش-کان-تاکر(KKT)[9] که برای تابع‌های محدب می‌باشد، استفاده می‌کنیم و برای نقاط بهینه مسئله دوگان کاربرد دارد که شرایط آن به‌صورت زیر است:

$$minf(x), \qquad s.t. \quad g_i(x) \leq \circ, \quad i = 1, \dots, m$$

با نوشتن تابع لاگرانژ آن داریم:

$$\ell(x, \alpha) = f(x) + \sum \alpha_i g_i(x)$$

برای کمینه کردن تابع لاگرانژ می‌توان از آن مشتق گرفت و برابر صفر قرار داد که در این صورت یک شرط جدید اضافه می‌شود:

$$\nabla \ell(x, \alpha)|_{x^*, \alpha^*} = \circ, s.t. \alpha_i^* \geq \circ, g_i(x^*) \leq \circ, \alpha_i^* g_i(x^*) = \circ,$$
$$i = 1, \dots, m$$

با توجه به مشابه بودن شرایط مسئله بهینه‌سازی ماشین بردار تکیه‌گاه با شرایط KKT و با در نظر گرفتن شرایط جدید می‌بایست از تابع لاگرانژ مشتق گرفت:

$$\nabla_W \ell(W, w_o, \alpha)|_{W^*, w_\circ^*, \alpha^*} = \circ \Rightarrow \frac{\partial \ell(W, w_\circ, \alpha)}{\partial w_\circ}|_{W^*, w_\circ^*, \alpha^*} = \circ,$$
$$\alpha_n^* > \circ$$

با توجه به شروط:

$$y_n(W^{*T} y + w_\circ^*) \geq 1, n = 1 \dots, N,$$
$$\alpha_i^* \underbrace{(1 - y_n(W^{*\top} y + w_\circ^*))}_{g_i(x^*)} = \circ$$

حاصل‌ضرب $\alpha_i^* \underbrace{(1 - y_n(W^{*\top} y + w_\circ^*))}_{g_i(x^*)} = \circ$ در دو زمان اتفاق می‌افتد، زمانی که α_i^* برابر صفر باشد آنگاه با توجه به شرایط $g_i(x^*) \leq \circ$، داده‌ها در دو طرف حاشیه قرار می‌گیرد، چون

$$g_i(x^*) \leq \circ, 1 - y_n(W^{*\top} y + w_\circ^*) \leq \circ, y_n(W^{*\top} y + w_\circ^*) \geq 1, \quad (۸)$$

اما اگر $g_i(x^*) = \circ$ و $\alpha > \circ$ باشند آنگاه می‌توان نوشت:

$$g_i(x^*) = \circ, 1 - y_n(W^{*\top} y + w_\circ^*) = \circ, y_n(W^{*\top} y + w_\circ^*) = 1, \quad (۹)$$

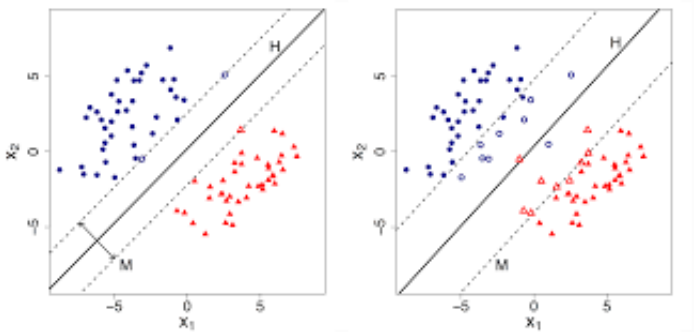
آنگاه داده‌ها روی حاشیه قرار می‌گیرند پس می‌توان W را به فرمی نوشت که فقط αهایی را در نظر بگیرد که مثبت هستند زیرا در بقیه قسمت‌ها صفر می‌شود یعنی درست دسته‌بندی‌شده است:

$$W = \sum \alpha_n y_n x_n = \sum_{\alpha > \circ} \alpha_n y_n x_n$$

به نقاطی که مقدار α آن‌ها بزرگ‌تر از صفر است، ماشین بردار تکیه‌گاه (SVM) می‌گوییم. با حل مسئله برنامه‌نویسی درجه دوم مقادیر $[\alpha_1, \dots, \alpha_p]$ محاسبه شد. حال داده‌هایی را که α_i آن‌ها بزرگ‌تر از صفر هستند، جدا کرده و مقدار $W = \sum \alpha_n y_n x_n$ محاسبه می‌شود.

$$|W^\top x_n + w_\circ| = 1 \Rightarrow y_n(W^\top x_n + w_\circ) = 1 \Rightarrow w_\circ = y_n - W^\top x_n$$

[9] Karush–Kuhn–Tucker



شکل ۱: نمایش داده‌ها و جدا شدن آن‌ها توسط حاشیه سخت (تصویر سمت چپ) و حاشیه نرم (تصویر سمت راست)

همچنین برای انتخاب مرز طبقه‌بندی‌ها داریم:

$$\hat{y} = sign(w_\circ + W^\top x)$$

$$\hat{y} = sign(y_i - \sum_{\alpha_n \geq \circ} \alpha_n y_n x_n^\top x + \sum_{\alpha_n \geq \circ} \alpha_n y_n x_n^\top x_n)$$

حاصل‌ضرب داخلی $nx_n^\top x_n$ را می‌توانیم با هسته جایگزین کرد تا وقتی به ابعاد دیگر می‌روند محاسبه ضرب داخلی طولانی نباشد. تمام مراحل فوق انجام شد تا با استفاده از داده‌های کمتر بتوان مرز را مشخص کرد. تمام مطالبی که گفته شد، مربوط به حاشیه سخت بود که در مورد داده‌هایی که در بین حاشیه بودند نظری نداشت و فقط داده‌هایی که روی حاشیه بودند، موردبررسی قرار گرفت.

۲.۲ مسئله بهینه‌سازی حاشیه نرم

هنگامی‌که داده‌ها وارد حاشیه شوند آنگاه با مفهوم جدیدی به نام حاشیه نرم [10] ماشین بردار تکیه‌گاه روبرو می‌شویم. تفاوت میان حاشیه سخت و نرم ماشین بردار تکیه‌گاه همان‌طور در شکل ۱ مشخص است، به‌گونه‌ای می‌باشد که در حاشیه سخت داده‌ای درون حاشیه قرار نمی‌گیرد، اما در حاشیه نرم این انعطاف‌پذیری وجود دارد که داده‌ها داخل حاشیه باشند.

اکنون برای حاشیه نرم می‌توان نوشت:

$$y_n(W^\top x_n + w_\circ) \geq 1 - \xi_n, \qquad \xi_n > \circ, \qquad (10)$$

در معادله‌ی بالا ξ_n فاصله هر داده تا خط حاشیه است، پس اگر داده‌ها درست دسته‌بندی شوند آنگاه $\xi_n = \circ$ می‌باشد.

اکنون یک مسئله بهینه‌سازی وجود دارد که داده‌ها امکان دارد داخل حاشیه قرار بگیرند. در حاشیه سخت تلاش برای کمینه کردن $\frac{1}{2}\|W\|$ بود ولی در اینجا علاوه بر انجام کار فوق می‌توان مجموع ξ_n ها را هم به حداقل رساند. پس برای بهینه‌سازی داریم:

$$\min \frac{1}{2}\|W\| + c\sum \xi_n, \quad s.t. \quad y_n(W^T x_n + w_\circ) \geq 1 - \xi_n,$$
$$n = 1, \ldots, N, \xi_n \geq \circ$$

با توجه به تابع لاگرانژ داریم:

$$\ell(W, w_\circ, \xi, \alpha, \beta) = \frac{1}{2}\|W\|^2 + C\sum_{n=1}^{N} \xi_n,$$
$$+ \sum_{n=1}^{N} \alpha_n(1 - \xi_n - y_n(W^T x_n + w_\circ)) - \sum_{n=1}^{N} \beta_n \xi_n, \qquad (11)$$

با توجه به آنچه گفته شد باید $W, w_\circ, \xi$ را کمینه و ضرایب لاگرانژ را بیشینه کرد. لذا با مشتق‌گیری و برابر صفر قرار دادن خواهیم داشت:

$$\nabla_W \ell(W, w_\circ, \xi, \alpha, \beta) = \circ \Rightarrow W = \sum_{n=1}^{N} \alpha_n y_n x_n,$$

$$\frac{\partial \ell(W, w_\circ, \xi, \alpha, \beta)}{\partial w_\circ} = \circ \Rightarrow -\sum_{n=1}^{N} \alpha_n y_n = \circ,$$

$$\frac{\partial \ell(W, w_\circ, \xi, \alpha, \beta)}{\partial \xi} = \circ \Rightarrow c - \alpha_n - \beta_n = \circ,$$

حال با جایگذاری در معادله‌ی (۱۱) داریم:

$$\max_x \{\sum_{n=1}^{N} \alpha_n - \sum\sum \alpha_n \alpha_m y_n y_m x_n^\top x_m\},$$

$$s.t. \quad \sum_{n=1}^{N} \alpha_n y_n = \circ, \circ \leq \alpha_n \leq c, n = 1, \ldots, N$$

با توجه به اینکه مسئله فوق یک مسئله برنامه‌نویسی درجه دوم می‌باشد با حل کردن و به دست آوردن α می‌توان W را محاسبه کرد. در حاشیه نرم اگر داده‌ها شرط $\circ \leq \alpha_n < c$ را داشته باشند آنگاه داده‌ها روی حاشیه قرار می‌گیرند که همان بردارهای تکیه‌گاه هستند زیرا با توجه به

[10] Soft Margin

شرایط مسئله:

$$\alpha_i g_i(x^*) = ۰, \quad \beta_n \xi_n = ۰, \quad c = \alpha_n + \beta_n$$

وقتی β_n مقدار داشته باشد آنگاه با برقراری شرایط فوق، باید $\xi_n = ۰$ باشد. حال اگر $\alpha_n = c$ باشد در این صورت با توجه به مطالب گفته‌شده، به‌ناچار $\beta_n = ۰$ می‌باشد و لذا $\xi > ۰$ است، که این شرایط معادل داده‌هایی است که داخل حاشیه شدند یا اشتباه دسته‌بندی شدند.

برای تعیین کردن مقدار مناسب برای این پارامتر، از اعتبارسنجی متقابل [11] استفاده می‌کنیم (به [۲] و [۱۵] مراجعه شود). اگر مقدار $c = ۰$ باشد، در عمل اجازه داده‌شده که داده‌ها در هر فاصله‌ای باشند یعنی تلاش ما بی‌فایده بوده، و اگر مقدار c بزرگ تنظیم شود یعنی برای همه نمونه‌ها شرایط حاشیه برقرار است و اگر $c = \infty$ باشد همان حاشیه سخت است، زیرا ξ ناچار است بین صفر و یک باشد و اگر یک عدد بزرگ هم در آن ضرب شود، حدودی نزدیک صفر می‌شود یعنی جمله $c\sum\xi$ اضافه می‌شود.

۳ رگرسیون بردار تکیه‌گاه

ماشین بردار تکیه‌گاه در طبقه‌بندی بسیار قوی هستند اما در رگرسیون شناخته‌شده نیستند. رگرسیون بردار تکیه‌گاه [12] یک حالت از ماشین بردار تکیه‌گاه است که به‌جای گرفتن مقادیر گسسته ۱− و ۱ در متغیرهای پاسخ، مقادیر پیوسته می‌گیرند.

در ماشین بردار تکیه‌گاه هر چه تعداد داده کمتری درون حاشیه قرار گیرد، خط جداکننده مناسب‌تر می‌باشد اما در رگرسیون بردار تکیه‌گاه با در نظر گرفتن حاشیه‌ها، هرچه تعداد داده بیشتری درون حاشیه قرار بگیرد، مدل مناسب‌تر می‌باشد. در این بخش، هدف ساخت مدل روی داده‌های $\{x_k, y_k\}_{k=۱}^{N}$ با استفاده از رگرسیون بردار تکیه‌گاه که مقدار متغیر پاسخ آن پیوسته است. در اکثر مدل‌های رگرسیون هدف به حداقل رساندن مجموع مربع خطاها است مانند حداقل توان‌های دوم معمولی که به شکل زیر می‌باشد:

$$\min \sum_{i=۱}^{n} (y_i - w_i x_i)^۲$$

رگرسیون‌های ستیغی، لاسو و شبکه الاستیک با اضافه کردن یک پارامتر تاوان [13] باهدف به حداقل رساندن پیچیدگی و یا کاهش تعداد ویژگی‌های مورداستفاده در مدل نهایی استفاده می‌شود، اما در روش رگرسیون بردار تکیه‌گاه خطا قرار نیست کمینه شود بلکه قرار است محدوده‌ای داشته باشد که در یافتن مقادیر پیش‌بینی بتواند منعطف‌تر باشد. پس این مدل به ما انعطاف‌پذیری می‌دهد تا تعیین کنیم چه مقدار خطا در مدل قابل‌قبول است. همان‌طور که هانگ و تسای [14] (۲۰۰۹) و پاتل و همکاران [15] (۲۰۱۵) به نتیجه رسیدند، استفاده از فرمول زیر مناسب‌تر از فرمول (۱۰) می‌باشد:

$$f(x, W) = W^\top x + b, \qquad (۱۲)$$

برای دستیابی به این هدف، یک خطای آستانه ϵ تعریف‌شده است تا در معادله‌ی زیر به کمترین مقدار برسد:

$$|y - f(x,W)|_\epsilon = \begin{cases} ۰, & |y - f(x,W)| \le \epsilon, \\ |y - f(x,W)| - \epsilon, & o.w, \end{cases} \qquad (۱۳)$$

اکنون با تعریف R و با توجه به $\|W\|$ داریم:

$$R = \frac{۱}{۲}\|W\|^۲ + c\left(\sum_{i=۱}^{N} |y_i - f(x_i, W)|_\epsilon\right)$$

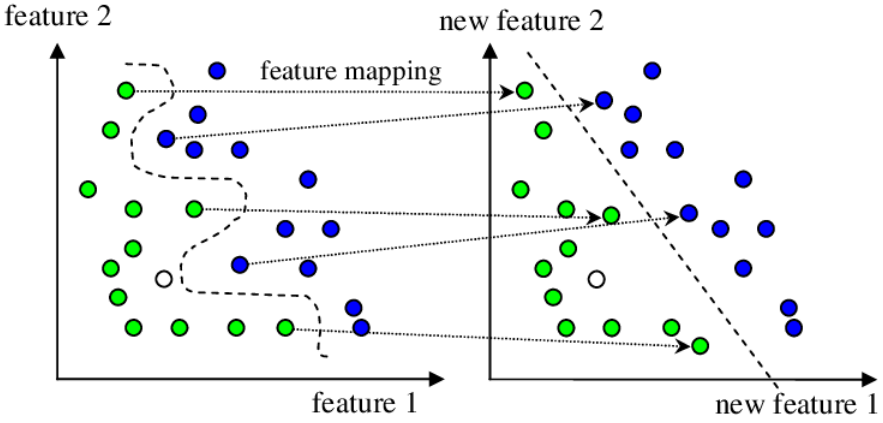
حال با تعریف متغیرهای ξ و ξ^* که فاصله هر داده تا خط حاشیه می‌باشد، معادله‌ی بالا تبدیل به معادله (۱۴) با شروط زیر می‌شود:

$$(W^\top x_i + b) - y_i \le \epsilon + \xi_i, y_i - (W^\top x_i + b) \le \epsilon + \xi_i^*,$$

$$\xi_i, \xi_i^* \ge ۰,$$

$$R = \frac{۱}{۲}\|W\|^۲ + c\sum_{i=۱}^{N}(\xi_i + \xi_i^*), \qquad (۱۴)$$

[11] Cross Validation
[12] Support Vector Regression
[13] Penalty Parameter
[14] Huang and Tsai
[15] Patel et al



شکل ۲: در این شکل با توجه به وجود مرز غیرخطی بین داده‌ها، ابتدا آن‌ها را به فضای جدید برده سپس می‌توان مرز خطی پیدا کرد.

۴ عدم وجود مرزخطی

همان‌طور که در شکل ۲ دیده می‌شود، اگر مرز خطی وجود نداشته باشد، باید داده‌ها را به فضای جدید برده و در آن فضا برای داده‌ها مرز خطی پیدا کنیم.

$$x \to \Phi\left(x = [\phi_1(x), \ldots, \phi_n(x)]\right)$$
$$R^d \to R^m$$

در همه‌ی مسائل فوق باید X را به $\Phi(x)$ تبدیل کنیم، اما چون داده‌ها همه وارد فضای جدید می‌شوند، لذا محاسبه‌ی ضرب داخلی $\Phi(x')\Phi(x)^\top$ بسیار طولانی است. بنابراین با معرفی روش زیر بدون اینکه داده‌ها را به فضای جدید ببریم، می‌توان ضرب داخلی را حساب کرد. یکی از این راه‌ها استفاده از حقه هسته [۱۶] است. چهار هسته معروف ماشین‌های یادگیری تکیه‌گاه عبارت‌اند از:

- هسته خطی [۱۷]: ساده‌ترین تابع هسته است که حاصل‌ضرب داخلی $< x, y >$ به‌علاوه یک مقدار ثابت اختیاری c می‌باشد.

$$k(x, y) = x^\top y + c$$

اگر با رسم داده‌ها متوجه رابطه خطی شدیم یا آنکه بتوان آن‌ها را با یک خط از هم جدا کرد، استفاده از این روش سودمند می‌باشد. در مجموع هسته خطی دقت کمتری نسبت به هسته‌های چندجمله‌ای و ... دارد.

- هسته چندجمله‌ای [۱۸]: هسته چندجمله‌ای برای زمانی که کلیه داده‌های آموزش نرمال شده‌اند، مناسب‌تر است. فرم هسته آن به‌صورت زیر می‌باشد:

$$k(x, y) = (\alpha x^\top y + c)^d$$

که در هسته بالا پارامترهای c، α و درجه‌ی چندجمله‌ای d قابل تنظیم است.

- هسته گوسی: نمونه‌ای از تابع شعاعی [۱۹] است که هسته آن به‌صورت زیر است:

$$k(x, y) = exp\left(\frac{\|x - y\|^2}{2\sigma^2}\right)$$

که پارامتر قابل تنظیم σ نقش عمده‌ای در تابع هسته دارد به‌طوری‌که اگر بیش‌ازحد بزرگ تعیین شود، نمایی رفتاری تقریباً خطی می‌گیرد و پیش‌بینی بر اساس آن شروع به از دست دادن رفتار غیرخطی خود خواهد کرد. اگر بیش‌ازحد کوچک تعیین گردد، تابع فاقد تنظیم و مرز تصمیم‌گیری در مورد اختلال در داده‌های آموزش بسیار حساس خواهد بود.

[16] Kernel Trick
[17] Linear
[18] Polynomial
[19] Radial
[20] Perceptron

- هسته سیگموید: به‌عنوان هسته چندلایه پرسپترون [۲۰] شناخته می‌شود. هسته سیگموید از قسمت شبکه‌ی عصبی می‌آید و زمانی که تابع سیگموئید دو قطبی است، اغلب به‌عنوان تابعی از فعال‌سازی نورون‌های مصنوعی استفاده می‌شود. هسته آن به‌صورت زیر می‌باشد:

$$k(x,y) = \tanh(\alpha x^\top y + c)$$

این هسته به دلیل منشأ تئوری شبکه عصبی برای ماشین بردار تکیه‌گاه کاملاً رواج دارد. همچنین، در عمل مشخص‌شده است که عملکرد خوبی دارد. دو پارامتر قابل تنظیم در هسته سیگموید وجود دارد که عبارت‌اند از شیب α و عرض از مبدأ c.

تکنیک رگرسیون بردار تکیه‌گاه برای ساخت مدل به توابع هسته متکی است، انتخاب کدام هسته با توجه به کدام داده مناسب‌تر است و کدام عملکرد بهتری دارد، بسیار دشوار است و نیاز به تکنیک بهینه‌سازی دارد.

۵ تحلیل داده‌های سهام

در این بخش، با بررسی یکی از سهام بازار بورس اوراق بهادر به دنبال یافتن بهترین مدل به‌منظور پیش‌بینی قیمت‌های سهام هستیم. با در نظر گرفتن سهام های‌وب در بازه زمانی ۴۳۲ روزه و داشتن اطلاعات قیمت ابتدایی، کمترین قیمت، بیشترین قیمت، ارزش معاملات و قیمت بسته شدن سهم در یک روز با استفاده از دو روش رگرسیون بردار تکیه‌گاه و رگرسیون نیم‌پارامتری، به دنبال یافتن مدل مناسب و پیش‌بینی سهام موردنظر هستیم. در این داده‌ها، قیمت پایانی را متغیر وابسته و سایر متغیرها را مستقل در نظر می‌گیریم.

۱.۵ ماشین بردار تکیه‌گاه

در این قسمت، با در نظر گرفتن مدل زیر، به دنبال یافتن بهترین برآورد برای ضرایب در داده‌های معرفی‌شده می‌باشیم:

$$Y = w_0 + \sum_{i=1}^{5} w_i X_i + \epsilon, \qquad (۱۵)$$

در مدل فوق Y قیمت بسته شدن سهم، w_i ضرایب مدل رگرسیون بردار تکیه‌گاه و X_i به ترتیب برابر تاریخ، حجم خریدوفروش، قیمت ابتدایی، کمترین قیمت و بیشترین قیمت می‌باشد. محاسبات مربوطه در نرم‌افزار R انجام‌شده است. برآورد ضرایب مدل (۱۵) با هسته‌های متفاوت را می‌توان در جدول (۱) مشاهده کرد. در جدول (۲) مدل‌های بسته‌شده روی داده‌ها بر اساس سه معیار مجذور همبستگی، ریشه میانگین توان‌های دوم خطا و میانگین قدرمطلق درصد خطا مقایسه شد. همان‌طور که دیده می‌شود، بر اساس نتایج به‌دست‌آمده در جدول (۲)، مدل رگرسیون بردار تکیه‌گاه با هسته گوسی با خطای برابر صفر بهترین نتیجه را نسبت به سایر هسته‌ها نشان می‌دهد.

جدول ۱: برآورد ضرایب در مدل(۱۵) به روش ماشین بردار تکیه‌گاه با هسته‌های متفاوت

برآوردگر	SVR(خطی)	SVR(چندجمله‌ای)	SVR(شعاعی)	SVR(سیگموید)
$\hat{w}_0$	−۰٫۰۴۸	۰٫۲۳	−۱٫۰۲	−۰٫۴۵
$\hat{w}_1$	۰٫۰۶۹	۲۴٫۱۷	۱۶٫۰۹	۱٫۳۳
$\hat{w}_2$	−۰٫۰۶۵	۶٫۳۷	−۱۱٫۰۶	−۱٫۷۲
$\hat{w}_3$	−۰٫۰۹	۲۷٫۷۰	۱۸٫۱۰	۴٫۸۰
$\hat{w}_4$	−۰٫۲۹	۲۷٫۹۳	۱۶٫۷۷	۴٫۷۹
$\hat{w}_5$	۱٫۳۱	۲۷٫۶۰	۲۰٫۸۹	۴٫۸۸

اکنون با تغییر مقدار خطا از صفر به ۰٫۱ در مدل گوسی به دنبال مدلی انعطاف‌پذیرتر هستیم. همان‌طور که در جدول (۳) آشکار است، مدل گوسی شعاعی با پارامترهای، $\gamma = ۰٫۲۵$ ، $c = ۱۰۰$ ، $\epsilon = ۰٫۱$ و تعداد بردارهای تکیه‌گاه ۲۷ نسبت به مدل گوسی با خطای صفر با توجه به افزایش مجذور همبستگی و کاهش ریشه میانگین توان‌های دوم خطا مدلی مناسب‌تر می‌باشد.

برآورد ضرایب مدل (۱۵) به روش بردار تکیه‌گاهی گوسی با خطای ۰٫۱ در جدول (۴) گزارش‌شده است.

جدول ۲: مقایسه ماشین بردار تکیه‌گاه با هسته‌های متفاوت در مدل (۱۵) بر اساس معیارهای برازش متفاوت

مدل	هسته	مجذور همبستگی	ریشه میانگین توان‌های دوم خطا	میانگین قدرمطلق درصد خطا
SVR	خطی	۰٫۸۲	۳۶۹۹٫۵۶۲	۰٫۱۹
SVR	چندجمله‌ای	۰٫۸۳	۳۶۲۹٫۴۴	۰٫۲۲
SVR	سیگموید	۰٫۵۵	۵۸۸۵٫۹۴۶	۰٫۵۱۳۳
SVR	گوسی	۰٫۹۵	۱۹۶۸٫۱۴۱	۰٫۰۱۸

جدول ۳: مقایسه دو مدل شعاعی با تغییر در مقدار خطا

	$\epsilon = ۰$	$\epsilon = ۰٫۱$
آزمون	مدل SVR با هسته شعاعی	مدل SVR با هسته شعاعی
مجذور همبستگی	۰٫۹۵	۰٫۹۷
ریشه میانگین توان‌های دوم خطا	۱۹۶۸٫۱۴۱	۱۴۵۱٫۳۲۹
میانگین قدرمطلق درصد خطا	۰٫۰۱۸	۰٫۰۵۵

جدول ۴: برآورد ضرایب مدل (۱۵) به روش بردار تکیه‌گاه گوسی با خطای ۰٫۱ (بهترین مدل)

برآورد ضرایب	ضرایب
$\hat{w}_۰$	−۱٫۲۲
$\hat{w}_۱$	۱۴٫۳۲
$\hat{w}_۲$	−۱۳٫۶۴
$\hat{w}_۳$	۱۲٫۵۵
$\hat{w}_۴$	۱۲٫۹۸
$\hat{w}_۵$	۱۵٫۳۱

۲.۵ مدل نیم‌پارامتری

برای برازش این مدل، ابتدا به دنبال تعیین رابطه پارامتری و ناپارامتری متغیرهای مستقل با متغیر وابسته بوده و سپس به برآورد ضرایب پارامتری و برازش تابع ناپارامتری می‌پردازیم (برای دیدن جزئیات بیشتر [۱۴] و [۱۶] را ببینید). در اینجا با استفاده از نمودار متغیر افزوده [21] بخش ناپارامتری مدل را تشخیص دهیم. نمودار متغیر افزوده به‌طور شهودی اثر هر یک از متغیرهای پیشگو را پس از حذف اثر سایر متغیرهای پیشگو، بر متغیر پاسخ آشکار می‌کند. با مشاهده به نمودار متغیر افزوده در شکل ۳ ازآنجاکه شواهدی مبنی بر رابطه جزئی میان متغیر پاسخ مصرف قیمت بسته شدن و متغیر قیمت ابتدایی $(PriceFirst)$ یک رابطه غیرخطی است از نمودار به دست می‌آید، این متغیر به‌عنوان بخش ناپارامتری مدل در نظر گرفته می‌شود، با تغییر متغیرهای زیر داریم:

$$Y = \beta_۰ + \sum_{i=۱}^{۴} \beta_i X_i + f(t) + \epsilon \qquad (۱۶)$$

[21] Added Variable Plot

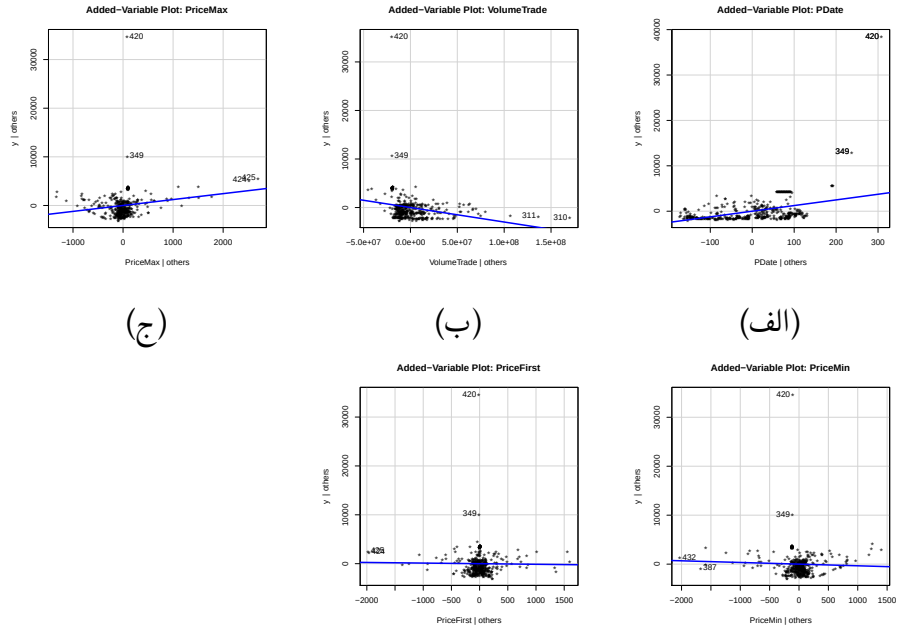
در مدل بالا Y قیمت بسته شدن سهم، β_i ضرایب قسمت پارامتری مدل نیم‌پارامتری، X_i به ترتیب برابر تاریخ، حجم خریدوفروش، کمترین قیمت و بیشترین قیمت و $f(t)$ قسمت ناپارامتری مدل نیم‌پارامتری می‌باشد که مربوط به داده‌های قیمت ابتدایی است.

برآورد ضرایب بر اساس مدل نیم‌پارامتری (۱۶) در جدول (۵) گزارش‌شده است.

اکنون می‌توان مدل نیم‌پارامتری برازش شده را توسط سه معیار مجذور همبستگی، ریشه میانگین توان‌های دوم خطا و میانگین قدرمطلق درصد خطا در جدول ۶ مورد ارزیابی و مقایسه با رگرسیون بردار تکیه‌گاه قرار داد. همچنین، برازش قسمت ناپارامتری (قیمت ابتدایی سهم) در نمودار ۴ نشان داده‌شده است.

۶ بحث و نتیجه‌گیری

ماشین بردار تکیه‌گاه که شاخه‌ای از یادگیری ماشین هستند، در مدل‌سازی بر روی داده‌ها دارای عملکرد بسیار مناسب بوده و با توجه به ویژگی تغییر در مقدار خطا، می‌توان مدل‌های منعطف‌تری را برازش نمود.

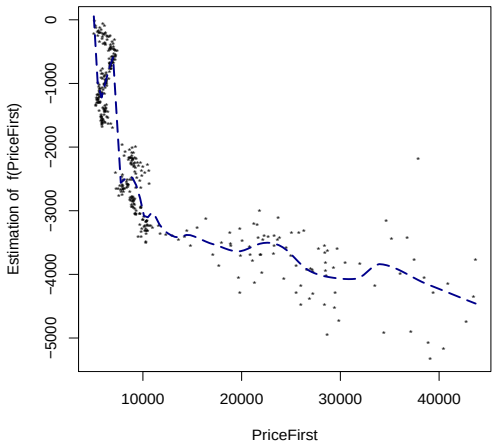


(الف) (ب) (ج)

(د) (ه)

شکل ۳: نمودار افزوده متغیرهای مستقل در مقابل متغیر وابسته، نمودارهای (الف)، (ب)، (ج) و (د) دارای ارتباط پارامتری و نمودار (ه) دارای ارتباط ناپارامتری با متغیر وابسته می‌باشد.

در این مقاله از مدل ماشین بردار تکیه‌گاه با چهار هسته متفاوت استفاده کردیم و مدل رگرسیون نیم‌پارامتری را بر روی یکی از سهام اوراق بورس بهادار (سهام های‌وب) پیاده‌سازی کردیم. بر اساس نتایج به‌دست‌آمده، مشخص شد که مدل رگرسیون بردار تکیه‌گاه با هسته شعاعی و خطای ۰/۱ بهترین برازش را در مقایسه با سایر مدل‌های رگرسیون بردار تکیه‌گاه با هسته‌های متفاوت داشته و همچنین، در مقایسه مدل مورداشاره با مدل نیم‌پارامتری کلاسیک، طبق نتایج جدول ۶، مشخص شد که مقدار مجذور همبستگی بالاتر و ریشه میانگین توان‌های دوم خطا و میانگین قدرمطلق درصد خطا پایین‌تری از مدل نیم‌پارامتری داشته و بنابراین کاراتر است.



شکل ۴: برازش تابع ناپارامتری

جدول ۵: برآورد ضرایب قسمت پارامتری

برآوردگر	مدل نیم‌پارامتری(قسمت پارامتری)
$\hat{\beta}_0$	۰
$\hat{\beta}_1$	۹٫۵۸۴۷۵۳
$\hat{\beta}_2$	$۱٫۶۷۵۲۴۷e^{-06}$
$\hat{\beta}_3$	$۵٫۴۳۷۴۸۰e^{-01}$
$\hat{\beta}_4$	$۴٫۶۱۷۷۵۲e^{-01}$

جدول ۶: مقایسه روش نیم‌پارامتری بر اساس معیارهای برازش

مدل	مجذور همبستگی	ریشه میانگین توان‌های دوم خطا	میانگین قدرمطلق درصد خطا
مدل نیم‌پارامتری	۰٫۹۰	۲۷۱۹٫۱۵۱	۰٫۶۳
مدل SVR با خطای ۰٫۱	۰٫۹۷	۱۴۵۱٫۳۲۹	۰٫۰۵۵

تقدیر و تشکر

نویسندگان مقاله ضمن تشکر از اعضای محترم هیئت تحریریه مجله، از پیشنهادها و نظرات ارزشمند داوران و ویراستار محترم مقاله که موجب ارتقاء سطح آن گردید کمال تشکر و قدردانی را دارند.

مراجع

[۱] روزبه، م. و امینی، م. (۱۳۹۸)، برآوردگر استوار مرزبندی شده تعمیم یافته محتمل در مدل رگرسیون نیمه پارامتری، *مجله علوم آماری*، ۱۳، ۴۴۱-۴۶۰.

[2] Amini, M. and Roozbeh, M. (2015), Optimal partial ridge estimation in restricted semiparametric regression models, *Journal of Multivariate Analysis*, **136**, 26-40.

[3] Araújo, R. D. A., Oliveira, A. L. and Meira, S. (2015), A hybrid model for high-frequency stock market forecasting, *Expert Systems with Applications*, **42**, 4081-4096.

[4] Blanchflower, D. G. and Oswald, A. J. (1994), *The wage curve*, MIT press.

[5] Choudhury, S., Ghosh, S., Bhattacharya, A., Fernandes, K. J. and Tiwari, M. K. (2014), A real time clustering and SVM based price-volatility prediction for optimal trading strategy, *Neurocomputing*, **131**, 419-426.

[6] Dash, R. and Dash, P. K. (2016), A hybrid stock trading framework integrating technical analysis with machine learning techniques, *The Journal of Finance and Data Science*, **2**, 42-57.

[7] Engle, R. F., Granger, C. W. J., Rice, J. and Weiss, A. (1986), Semiparametric Estimates of the Relation Between Weather and Electricity Sales, *Journal of the American Statistical Association*, **81**, 310-320.

[8] Fama, E. (1970), Efficient capital markets: A review of theory and empirical work, *The journal of Finance*, **25**, 383-417.

[9] Gerlein, E. A., McGinnity, M., Belatreche, A. and Coleman, S. (2016), Evaluating machine learning classification for financial trading: An empirical approach, *Expert Systems with Applications*, **54**, 193-207.

[10] Kao, L. J., Chiu, C. C., Lu, C. J., and Yang, J. L. (2013), Integration of nonlinear independent component analysis and support vector regression for stock price forecasting, *Neurocomputing*, **99**, 534-542.

[11] Manahov, V., Hudson, R. and Gebka, B. (2014), Does high frequency trading affect technical analysis and market efficiency and if so, how *Journal of International Financial Markets, Institutions and Money*, **28**, 131-157.

[12] Nayak, R. K., Mishra, D. and Rath, A. K. (2015), A naïve svm-knn based stock market trend reversal analysis for indian benchmark indices, *Applied Soft Computing*, **35**, 670-680.

[13] Patel, J., Shah, S., Thakkar, P. and Kotecha, K. (2015), Predicting stock market index using fusion of machine learning techniques, *Expert Systems with Applications*, **42**, 2162-2172.

[14] Roozbeh, M. (2015), Shrinkage ridge estimators in semiparametric regression models, *Journal of Multivariate Analysis*. **136**, 56-74.

[15] Roozbeh, M. (2018), Optimal QR-based estimation in partially linear regression models with correlated errors using GCV criterion, 117, *Computational Statistics & Data Analysis*. **117**, 45-61.

[16] Roozbeh, M. and Arashi, M. (2013), Feasible ridge estimator in partially linear models, *Journal of Multivariate Analysis*. **116**, 35-44.

[17] Vapnik, V. N. (1995), The Nature of Statistical Learning Theory, New York.

[18] Willis, R. J. (1986), Wage determinants: A survey and reinterpretation of human capital earnings functions. *Handbook of labor economics*, **1**, 525-602.

[19] Xiao, Y., Xiao, J., Lu, F. and Wang S. (2014), Ensemble anns-pso-ga approach for day-ahead stock e-exchange prices forecasting, *International Journal of Computational Intelligence Systems*, **7**, 272-290.

Support Vector Machines Regression Model and Comparison with Semi-parametric Regression

Mahdi Roozbeh[1] Arta Rouhi[2] Fatemeh Jahadi[3] Saeed Zalzadeh[4]

Abstract:

In this research, the aim is to assess and analyze a method to predict the stock market. However, it is not easy to predict the capital market due to its high dependence on politics but by data modeling, it will be somewhat possible to predict the stock market in the long period of time. In this regard, by using the semi-parametric regression models and support vector regression (SVR) with different kernels and measuring the predictor errors in the stock market of one stock based on daily fluctuations and comparing methods using the root of mean squared error and mean absolute percentage error criteria, support vector regression model has been the most appropriate fit to the real stock market data with radial kernel and error equal to 0.1.

Keywords: Machine Learning, Semi-parametric Regression Model, Stock Forecasting, Support Vector Regression Model.

[1]**Faculty of mathematics, Semnan university, Semnan, Iran.**

[2]**Master's degree graduate, statistics and Computer science, Semnan university, Semnan, Iran.**

[3]**Master's degree graduate, statistics and Computer science, Semnan university, Semnan, Iran.**

[4]**Faculty of mathematics, Semnan university, Semnan, Iran.**