Hindawi
Complexity
Volume 2022, Article ID 4740173, 13 pages
https://doi.org/10.1155/2022/4740173

*Research Article*

# Novel Feature Selection Method for Nonlinear Support Vector Regression

**Kejia Xu,[1] Ying Xu,[1] Yafen Ye ,[1] and Weijie Chen[2]**

[1]*School of Economics, Zhejiang University of Technology, Hangzhou 310023, China*
[2]*Zhijiang College, Zhejiang University of Technology, Hangzhou 310023, China*

Correspondence should be addressed to Yafen Ye; yafenye@163.com

Received 20 August 2022; Revised 3 November 2022; Accepted 12 November 2022; Published 30 November 2022

Academic Editor: Xiaoan Yan

The development of sparse techniques presents a major challenge to complex nonlinear high-dimensional data. In this paper, we propose a novel feature selection method for nonlinear support vector regression, called FS-NSVR, which first attempts to solve the nonlinear feature selection problem in the regression technology field. FS-NSVR preserves the representative features selected in the complex nonlinear system due to its use of a feature selection matrix in the original space. FS-NSVR is a challenging mixed-integer programming problem that is solved efficiently by using an alternate iterative greedy algorithm. Experimental results on three artificial datasets and five real-world datasets confirm that FS-NSVR effectively selects representative features and discards redundant features in a nonlinear system. FS-NSVR outperforms $L_1$-norm support vector regression, $L_1$-norm least squares support vector regression, and $L_p$-norm support vector regression on both feature selection ability and regression efficiency.

## 1. Introduction

High-dimensional data have commonly emerged in diverse fields, such as finance [1], economics [2], biology [3], and medicine [4]. Complex nonlinear relationships between features may exist in high-dimensional datasets [5]. For example, most economic and financial time series follow nonlinear behavior [6]. Ling et al. explored the nonlinear relationship between globalization, natural resources, financial development, and carbon emissions [6]. Another example of complex nonlinear high-dimensional data is in the medicine field. Medical costs have a sophisticated relation with features [7]. Complex nonlinear high-dimensional data have frequently emerged in the biology field. Nonlinear relations between features can depict biological relationships more precisely and reflect critical patterns in biological systems [8].

Complex nonlinear high-dimensional data may include some irrelevant and redundant features, which may reduce the effectiveness of data mining and may detract from the quality of the results [9–11]. Thus, complex nonlinear high-dimensional data need a sparse technique. Feature selection, as a useful sparse technique, selects some useful features upon which to focus its attention and ignores the rest [12–17]. In general, feature selection methods are classified as filter, wrapper, and embedded methods [16–18]. The embedded method is very popular for feature selection since it conducts feature selection and other learning tasks simultaneously [19].

Sparse support vector regression, as a branch of sparse support vector machine [20–23], is a computationally powerful feature selection method. Sparse support vector regression always adopts a sparse regularization term to realize feature selection and regression simultaneously. Therefore, sparse support vector regression is an embedded feature selection method [24–26]. $L_1$-norm support vector regression ($L_1$-SVR) [27] and $L_1$-norm least squares support vector regression ($L_1$-LSSVR) [28] use the $L_1$-norm sparse regularization term to shrink some coefficients of the regression estimators towards 0. According to the regression estimators, the contribution of each feature to the final decision function can be judged, and then, the useful features are selected, while the irrelevant and redundant features are discarded. To improve the sparseness of $L_1$-SVR, Zhang et al.

[29] proposed $L_p$-norm support vector regression ($L_p$-SVR) ($0 < p < 1$). The $L_p$-norm regularization term in $L_p$-SVR shrinks more coefficients of the regression estimators towards 0, and some coefficients are shrunk to exactly 0, leading to more irrelevant and redundant features being discarded. However, $L_1$-SVR, $L_1$-LSSVR, and $L_p$-SVR only solve the linear feature selection problem, which is not always suitable for complex nonlinear cases.

To solve the feature selection problem for complex nonlinear high-dimensional data, we follow the spirit of nonlinear support vector machine-based feature selection [9, 30, 31] and then propose a novel feature selection method for nonlinear support vector regression, which is called FS-NSVR. We bring a feature selection matrix, as a diagonal matrix with elements of either 1 or 0, into nonlinear support vector regression. As a result, FS-NSVR becomes a mixed-integer programming problem (MPP). To solve FS-NSVR efficiently, we employ an alternate iterative greedy algorithm to find a local optimal value [32], in which we iteratively solve the standard SVR problem and a smaller nonconvex feature selection problem. In addition, a feature-ranking strategy is suggested [33], which ranks the features according to their contributions to the objective in the MPP. Compared with $L_1$-SVR, $L_1$-LSSVR, and $L_p$-SVR, the experimental results show that FS-NSVR selects the most appropriate representative features in the highly complex nonlinear relationships with smaller estimation errors than those produced by $L_1$-SVR, $L_1$-LSSVR, and $L_p$-SVR. This means that FS-NSVR not only selects the representative features but also has good regression effectiveness. The contributions of this paper are summarized as follows:

(1) Bringing a feature selection matrix into nonlinear support vector regression, we propose a novel feature selection method for nonlinear support vector regression to identify complex nonlinear relationships between features in their original space. The proposed model first attempts to solve the nonlinear feature selection problem in the regression technology field.

(2) The proposed model is a complex mixed-integer programming problem. Ensuring the efficiency of the learning process, we employ an alternate iterative greedy algorithm to find a local optimal value for the proposed model. The alternate iterative greedy algorithm, transferring the complex mixed-integer problem into the min-max optimization problem, effectively reduces the computational complexity.

(3) Experimental results on both artificial and real-world datasets indicate that the proposed model preserves the representative features selected in the complex nonlinear system, outperforming the other three linear feature selection methods, with better feature selection and regression results. The training speed of the proposed method confirms the efficiency of the alternate iterative greedy algorithm.

The remainder of this paper is organized as follows: Section 2 in this paper briefly focuses on support vector regression. In Section 3, we propose feature selection for nonlinear support vector regression. Section 4 provides artificial and real-world dataset experiments, and Section 5 concludes the paper.

## 2. Background

Starting with the notations, we consider a regression problem in $n$-dimensional real vector space $R^n$. Suppose that $Y = (y_1, \ldots, y_l)'$ is the response vector, $A$ is a known $l \times n$ design matrix of covariates, and $A_i = (x_{i1}, \ldots x_{in})'$ is the $n$ -dimensional training sample. Next, we briefly review support vector regression (SVR) [26] that are closely related to FS-NSVR.

The optimal nonlinear regression function of SVR is constructed as follows:

$$f(x) = K\left(x^T, A^T\right)w + b, \tag{1}$$

where $w \in R^n$, $b \in R$, and $K$ are an appropriately chosen kernel. Parameters in function (1) are estimated by solving the following optimization problem:

$$\begin{aligned} &\min_{w,b,\xi,\xi^*} \frac{1}{2}\|w\|_2^2 + Ce^T(\xi + \xi^*), \\ &\text{subjected.to. } Y - K\left(A, A^T\right)w - b \le \varepsilon + \xi, \xi \ge 0, \\ &K\left(A, A^T\right)w + b - Y \le \varepsilon + \xi^*, \xi^* \ge 0. \end{aligned} \tag{2}$$

where $\xi$ and $\xi^*$ are the slack variables and $C(C > 0)$ is a parameter determining the trade-off between the empirical risk and the regularization term. To derive the dual formulation of SVR, we first introduce the Lagrangian function for the problem (2), which is

$$\begin{aligned} L = &\frac{1}{2}w^T w + Ce^T(\xi + \xi^*) - \alpha^T\left(\varepsilon + \xi - Y + K\left(A, A^T\right)w + b\right) \\ &- \beta^T\left(\varepsilon + \xi^* + Y - K\left(A, A^T\right)w - b\right) - \eta^T\xi - \gamma^T\xi^*, \end{aligned} \tag{3}$$

where $\alpha, \beta, \eta$, and $\gamma$ are the Lagrangian multiplier vectors. The Karush–Kuhn–Tucher (KKT) necessary and sufficient optimality conditions for the problem (2) are given by

$$\begin{aligned} w - K\left(A, A^T\right)^T\alpha + K\left(A, A^T\right)^T\beta &= 0, \\ -e^T\alpha + e^T\beta &= 0, \\ C - \alpha_i - \eta_i &= 0, \quad i = 1, \ldots, l, \\ C - \beta_i - \gamma_i &= 0, \quad i = 1, \ldots, l, \\ Y - K\left(A, A^T\right)w - b \le \varepsilon + \xi, &\quad \xi \ge 0, \\ K\left(A, A^T\right)w + b - Y \le \varepsilon + \xi^*, &\quad \xi^* \ge 0, \\ \alpha^T\left(\varepsilon + \xi - Y + K\left(A, A^T\right)w + b\right) &= 0, \\ \beta^T\left(\varepsilon + \xi^* + Y - K\left(A, A^T\right)w - b\right) &= 0, \\ \eta^T\xi = 0, &\quad \eta \ge 0, \\ \gamma^T\xi^* = 0, &\quad \gamma \ge 0. \end{aligned} \tag{4}$$

According to the previously mentioned KKT conditions, we obtain the dual formulation of problem (2) as follows:

$$\max_{\alpha,\beta} -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\left(\alpha_i-\beta_j\right)K\left(x_i,x_j\right)\left(\alpha_i-\beta_j\right)+\sum_{i=1}^{l}\left(\alpha_i-\beta_j\right)y_i+\sum_{i=1}^{l}\varepsilon\left(\alpha_i+\beta_i\right)\text{subjecte.to.}\sum_{\mathrm{i}=1}^{1}\left(\alpha_i-\beta_j\right)=00\le\alpha_i\le C,\quad i=1,2,\ldots,l,0\le\beta_{\mathrm{i}}\le C,\quad i=1,2,\ldots,l. \tag{5}$$

$w$ can be obtained from the solution $\alpha$ and $\beta$ of (5) by

$$w=K\left(A,A^T\right)^T(\alpha-\beta). \tag{6}$$

For any solution to (5), $\overline{\alpha}=(\overline{\alpha}_1,\overline{\alpha}_2,\ldots,\overline{\alpha}_l)^T$ and $\overline{\beta}=(\overline{\beta}_1,\overline{\beta}_2,\ldots,\overline{\beta}_l)^T$, if $\overline{\alpha}\neq 0$, the solution $\overline{b}$ to the problem (2) can be obtained in the following way:

(1) For any nonzero component $\overline{\alpha}_j\in(0,C)$,

$$\overline{b}=y_j-\sum_{i=1}^{l}\left(\overline{\alpha}_i-\overline{\beta}_i\right)K\left(x_i,x_j\right)+\varepsilon. \tag{7}$$

(2) For any nonzero component $\overline{\beta}_k\in(0,C)$,

$$\overline{b}=y_k-\sum_{i=1}^{l}\left(\overline{\alpha}_i-\overline{\beta}_i\right)K\left(x_i,x_k\right)-\varepsilon. \tag{8}$$

The final decision function can be constructed as

$$f(x)=\sum_{i=1}^{l}\left(\overline{\alpha}_i-\overline{\beta}_i\right)K\left(x_i,x\right)+\overline{b}. \tag{9}$$

## 3. Feature Selection for Nonlinear Support Vector Regression

*3.1. Problem Formulation.* In this section, we propose feature selection for nonlinear support vector regression. $Z=\operatorname{diag}(1\text{or}0)$ is an $n\times n$ feature selection matrix. We consider the following nonlinear regression function:

$$f(x)=K\left(x^TZ,ZA^T\right)w+b, \tag{10}$$

where $w\in R^n$, $b\in R$, and $K$ is an appropriately chosen kernel. $w$ and $b$ are the unknown parameters that need to be estimated. The optimal feature selection matrix $Z$ also needs to be searched simultaneously.

The estimator of the regression function (10) can be defined as the solution to the FS-NSVR optimization problem:

$$\begin{aligned}&\min_{w,b,\xi,\xi^*,Z}\ \frac{1}{2}\|w\|_2^2+Ce^T\left(\xi+\xi^*\right)+e^TZe\\&\text{s.t. } Y-K\left(AZ,ZA^T\right)w-be\le\varepsilon e+\xi,\quad \xi\ge 0,\\&K\left(AZ,ZA^T\right)w+be-Y\le\varepsilon e+\xi^*,\quad \xi^*\ge 0,\\&Z=\operatorname{diag}(1\text{or}0),\end{aligned} \tag{11}$$

where $\xi$ and $\xi^*$ are slack variables and $C>0$ is a parameter determining the trade-off between the empirical risk and the regularization term. In fact, the feature selection matrix $Z$ defines a subspace spanned by the selected features. Minimizing the term $e^TZe$ in the objective function (11) has the beneficial effect of suppressing variables to produce a sparse set of nonzero feature weights. Therefore, FS-NSVR has nonlinear feature selection ability.

*3.2. Problem Solution.* Obviously, the FS-NSVR optimization problem is a mixed-integer programming problem. We reformulate problem (11) as follows:

$$\begin{aligned}&\min_{Z}\min_{w,b,\xi,\xi^*}\ \frac{1}{2}\|w\|_2^2+Ce^T\left(\xi+\xi^*\right)+e^TZe\\&\text{s.t. } Y-K\left(AZ,ZA^T\right)w-be\le\varepsilon e+\xi,\quad \xi\ge 0,\\&K\left(AZ,ZA^T\right)w+be-Y\le\varepsilon e+\xi^*,\quad \xi^*\ge 0,\\&Z=\operatorname{diag}(1\text{or}0).\end{aligned} \tag{12}$$

Solving problem (12) to obtain global optimality is highly challenging and impractical [24]. We employ an alternate iterative greedy algorithm to find a local optimal value. First, we fix the integer part $Z$ and then obtain the solution to (12), which leads to solving the problem in the same manner as SVR. Similar to the deduction process of nonlinear SVR in Section 2, we obtain the dual formulation of the inner minimization problem. Then, problem (12) can be rewritten as

$$\begin{aligned}
&\min_{Z}\max_{\alpha,\beta} -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i-\beta_i)K\left(x_i\circ Z_{ii}, x_j\circ Z_{jj}\right)\left(\alpha_j-\beta_j\right)+\sum_{i=1}^{l}(\alpha_i-\beta_i)y_i-\varepsilon\sum_{i=1}^{l}(\alpha_i+\beta_i)+\sum_{k=1}^{n}Z_{kk}\\
&\text{subjecte.to.} \sum_{i=1}^{l}(\alpha_i-\beta_i)=0,\\
&0\le\alpha_i\le C,\quad i=1,2,\cdots,l,\\
&0\le\beta_i\le C,\quad i=1,2,\cdots,l,\\
&Z_{kk}\in\{0,1\},\quad k=1,\cdots,n.
\end{aligned}\tag{13}$$

Obviously, problem (13) is a challenging min-max optimization problem. Fixing the optimal solution of the inner maximization problem $(\alpha,\beta)$, we obtain the outer minimization integer problem, which leads to exhaustive computation of the objective for the possible $Z$.

To make the greedy algorithm work, we follow the strategy in [33] to initialize $Z$ to make the algorithm more stable. The value of each feature is computed after solving SVR by

$$\mathrm{Value}_{\text{feature}}(i)=\left\|\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i-\beta_i)K\left(x_iZ_{ii}, x_jZ_{jj}\right)\left(\alpha_j-\beta_j\right)\right\|. \tag{14}$$

The score of the $i$ th feature that reflects the importance among all the features is computed by

$$\mathrm{Score}_{\text{feature}}(i)=\frac{\mathrm{Value}_{\text{feature}}(i)}{\sum_j \mathrm{Value}_{\text{feature}}(j)}. \tag{15}$$

The greedy algorithm starts from an initial $Z^{(0)}$ generated by (15). If $\mathrm{Score}_{\text{feature}}$ is less than $1/n$, then $Z_{ii}^{(0)}=0$; otherwise, $Z_{ii}^{(0)}=1$. We then fix $Z^{(0)}$ and solve problem (13) to obtain $(\alpha^{(0)},\beta^{(0)})$. We calculate $\mathrm{Value}_{\text{feature}}$ and $\mathrm{Score}_{\text{feature}}$according to (14) and (15), respectively. $Z^{(1)}$ is updated by replacing $Z^{(0)}$ if objective (13) decreases more than the tolerance. After updating $Z^{(1)}$, $(\alpha^{(1)},\beta^{(1)})$ can be obtained again. The algorithm will be terminated if objective (13) decreases less than the tolerance. We summarize the procedures of the greedy approach in Algorithm 1 to give the feature selection method for nonlinear support vector regression. The proof of convergence of the greedy approach in Algorithm 1 can be obtained from Mangasarian and Kou [32].

Obtaining the solution of problem (13), $w$ can be obtained by

$$w=K\left(AZ, ZA^T\right)^T(\alpha-\beta). \tag{16}$$

For any solution to (15) $\overline{\alpha}=(\overline{\alpha}_1,\overline{\alpha}_2,\cdots,\overline{\alpha}_l)^T$ and $\overline{\beta}=(\overline{\beta}_1,\overline{\beta}_2,\cdots,\overline{\beta}_l)^T$, if $\overline{\alpha}\neq 0$, the solution $\overline{b}$ to problem (6) can be obtained in the following way:

(1) For any nonzero component $\overline{\alpha}_j\in(0,C)$,

$$\overline{b}=y_j-\sum_{i=1}^{l}\left(\overline{\alpha}_i-\overline{\beta}_i\right)K\left(x_i\circ Z_{ii}, x_j\circ Z_{jj}\right)+\varepsilon. \tag{17}$$

(2) For any nonzero component $\overline{\beta}_k\in(0,C)$,

$$\overline{b}=y_k-\sum_{i=1}^{l}\left(\overline{\alpha}_i-\overline{\beta}_i\right)K\left(x_i\circ Z_{ii}, x_k\circ Z_{kk}\right)-\varepsilon. \tag{18}$$

The final decision function can be constructed as

$$f(x)=\sum_{i=1}^{l}\left(\overline{\alpha}_i-\overline{\beta}_i\right)K\left(x_iZ_{ii}, x^TZ\right)+\overline{b}. \tag{19}$$

*3.3. Computational Complexity.* Concerning the computational complexity of FS-NSVR, we find that FS-NSVR includes two parts: one is repeatedly computing the inner maximization problem of (13), and the other is repeatedly computing (14) and (15). The first part requires solving one quadratic programming problem. The time complexity of this part is approximately $O(2l)^3$. For the second part, it is easy to compute with the fixed $(\alpha,\beta)$, and the computational complexity of this part is not more than $k$ times.

## 4. Experimental Results

To test the feature selection and regression effectiveness of the proposed FS-NSVR, we compare it with $L_1$-SVR [27], $L_1$-LSSVR [28], and $L_p$-SVR [29] by using three artificial datasets and seven real-world datasets. $L_1$-SVR, $L_p$-SVR, and $L_1$-LSSVR are embedded linear feature selection methods. All of these methods are implemented in the MATLAB R2019b environment on a PC running the 64-bit Windows XP OS on a 1.6 GHz Intel (R) processor with 16 GB of RAM.

For the feature selection of nonlinear support vector regression, we employ a Gaussian kernel, and its kernel parameter $p$ is selected from the set $\{2^{-8},\ldots,2^{8}\}$. Parameter $C$ is also selected from the set $\{2^{-8},\ldots,2^{8}\}$. The insensitive parameter $\varepsilon$ is fixed at 0.01. The optimal values of the parameters in the experiments are obtained by using the grid search method.

Let $m$ be the number of samples, $y_i^t$ is the $i$th test sample, $\widehat{y}_i^t$ is the predicted value of $y_i^t$, and $\overline{y}=1/m\sum_{i=1}^{m}y_i^t$ is the average value of $y_1^t,\ldots,y_m^t$. We use the following evaluation criteria to evaluate the variable selection and regression results.

```
Input: Training set $(A, Y)$. The appropriate kernel parameter $\delta$, parameter $C$;
Output: $\alpha$, $\beta$, and $Z$;
Begin
  Start with $Z^0 = I$; set iteration number $k = 1$;
  while$k < N$do
    Find the solution $(\alpha^{(k)}, \beta^{(k)})$ to problem (15) with the fixed $Z^{(k-1)}$; Compute each feature score by (15).
    for $i = 1 : n$
      If$\mathrm{Score}_{\mathrm{feature}}(i) < 1/n$then
$Z_{ii}^{(k)} = 0$
      else
$Z_{ii}^{(k)} = 1$
      end
    end
    If$\|\alpha^{(k+1)} - \beta^{(k+1)}\|^2 \approx \|\alpha^{(k)} - \beta^{(k)}\|^2$then
      Converged and output $\alpha^{(k+1)}, \beta^{(k+1)}$ and $Z^{(k)}$ as the final solutions.
    else
      set $k = k + 1$
    end
  end
Output $\alpha^{(k+1)}, \beta^{(k+1)}$ and $Z^{(k)}$ as the final solutions;
end
```

Algorithm 1: Feature selection for nonlinear support vector regression.

$P_1$: the proportion of simulation runs with nonzero coefficients are selected

$R^2$: the coefficient of determination $R^2$ is defined as

$$R^2 = \frac{\sum_{i=1}^{m} \left(\hat{y}_i^t - \overline{y}\right)^2}{\sum_{i=1}^{m} \left(y_i^t - \overline{y}\right)^2}. \tag{20}$$

NMSE: the normalized mean squared error (NMSE) is defined as

$$\mathrm{NMSE} = \frac{\sum_{i=1}^{m} \left(y_i^t - \hat{y}_i^t\right)^2}{\sum_{i=1}^{m} \left(y_i^t - \overline{y}\right)^2}. \tag{21}$$

RMSE: the root mean square error (RMSE) is defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^{m} \left(y_i^t - \hat{y}_i^t\right)^2}. \tag{22}$$

Thus, the smaller the values of NMSE and RMSE are, the more statistical information is captured from the selected variables.

*4.1. Artificial Datasets.* To test the nonlinear feature selection performance of FS-NSVR, we provide three artificial datasets. Specifically, we set $n = 200$, $l = 500$, and $\varsigma \sim N(0, 0.01)$. The first regression model is generated as follows:

$$\text{Type A: } y = \frac{\sin\left(\sqrt{x_{50}^2 + x_{100}^2 + x_{150}^2 + x_{200}^2 + x_{250}^2 + x_{300}^2 + x_{350}^2 + x_{400}^2 + x_{450}^2 + x_{500}^2}\right)}{\sqrt{x_{50}^2 + x_{100}^2 + x_{150}^2 + x_{200}^2 + x_{250}^2 + x_{300}^2 + x_{350}^2 + x_{400}^2 + x_{450}^2 + x_{500}^2}} + \varsigma, \tag{23}$$

where $x_i \in [-5\pi, 5\pi], i \in \{50, 100, 150, 200, 250, 300, 350, 400, 450, 500\}$.

The second regression function is as follows:

$$\begin{aligned} &\text{Type B: } y = \left(1 - \sqrt{x^2} + 2x^2\right) e^{-0.5x^2} + \varsigma, \\ &x^2 = x_{50}^2 + x_{100}^2 + x_{150}^2 + x_{200}^2 + x_{250}^2 + x_{300}^2 + x_{350}^2 + x_{400}^2 + x_{450}^2 + x_{500}^2, \end{aligned} \tag{24}$$

where $x_i \in [-10, 10], i \in \{50, 100, 150, 200, 250, 300, 350, 400, 450, 500\}$.

The third regression function is as follows:

$$\begin{aligned} &\text{Type C: } y = 6\sin(0.5\pi - x) + 3(\sin(0.5\pi - x))\varsigma, \\ &x = \sqrt{x_{50}^2 + x_{100}^2 + x_{150}^2 + x_{200}^2 + x_{250}^2 + x_{300}^2 + x_{350}^2 + x_{400}^2 + x_{450}^2 + x_{500}^2}, \end{aligned} \tag{25}$$

where $x_i \in [-10, 10], i \in \{50, 100, 150, 200, 250, 300, 350, 400, 450, 500\}$. The specifications of these artificial datasets are listed in Table 1.

To evaluate the performance of the feature selection results, we adopt the following criteria:

$$\begin{aligned} \text{Precision} &= \frac{tp}{tp + fp}, \\ \text{Recall} &= \frac{tp}{tp + fn}, \end{aligned} \tag{26}$$

where $tp$ are true positives, $fp$ are false-positives, $fn$ are false negatives, and $tn$ are negative counts. Precision and recall are commonly used to present results for binary decision problems in machine learning since they give a more accurate evaluation of an algorithm's performance. Here, we use precision and recall to evaluate the feature selection results of FS-NSVR, $\text{L}_1$-SVR, $\text{L}_\text{p}$-SVR, and $\text{L}_1$-LSSVR.

The best parameters of FS-NSVR, $\text{L}_1$-SVR, $\text{L}_\text{p}$-SVR, and $\text{L}_1$-LSSVR for artificial datasets are shown in Table 2. The feature selection and regression results on the previous three artificial datasets are shown in Table 3. From Table 3, we find that FS-NSVR drives larger precision and recall than $\text{L}_1$-SVR, $\text{L}_\text{p}$-SVR, and $\text{L}_1$-LSSVR. Meanwhile, FS-NSVR obtains a larger $R^2$ and smaller NMSE than $\text{L}_1$-SVR, $\text{L}_\text{p}$-SVR, and $\text{L}_1$-LSSVR. It is clear that FS-NSVR has the ability to select the representative features and discard the redundant features in the nonlinear system. Therefore, FS-NSVR is suitable for solving the nonlinear feature selection problem, while $\text{L}_1$-SVR, $\text{L}_\text{p}$-SVR, and $\text{L}_1$-LSSVR are not suitable for solving the feature selection problem of complex nonlinear high-dimensional data. In terms of running times, although the training speed of FS-NSVR is slower than $\text{L}_1$-LSSVR, it is significantly faster than $\text{L}_1$-SVR and $\text{L}_\text{p}$-SVR.

*4.2. Parameters and Nonlinear Feature Selection Analysis.* In this part, we analyze the effects of the parameters $p$ and C on the nonlinear feature selection results. To test the influence of the kernel parameter $p$ on NMSE, $R^2$, precision, and recall, we first fix parameter $C$ as the optimal value used in the experiments on the artificial datasets. Figures 1–3 illustrate the influence of the kernel parameter $p$ on the nonlinear feature selection results for Type A, Type B, and Type C, respectively. From Figures 1–3, we find that when $p$ increases, the NMSE value decreases and then increases. As the kernel parameter $p$ increases, $R^2$ and precision increase and then decrease, which means that the kernel parameter $p$ has a strong influence on the feature selection ability of FS-NSVR. When $p$ is selected as the optimal value, precision and recall reach maximum values, which means that FS-NSVR selects representative features and discards irrelevant features. Although the number of selected features is small, FS-NSVR can select the representative features in the datasets.

To further test the influence of parameter $C$ on the FS-NSVR feature selection results, we fix the kernel parameter $p$ as the optimal value used in the experiments on the artificial datasets. Figures 4–6 show the influence of $C$ on NMSE, $R^2$, precision, and recall. From Figures 4–6, we observe that as parameter $C$ increases, NMSE decreases and then remains constant. When $C = 2$, precision reaches a maximum value and then remains constant, which means that FS-NSVR selects the representative features and discards the irrelevant features.

*4.3. Real-World Datasets.* To further test the feature selection and regression performance of FS-NSVR, we consider five real-world datasets from the UC Irvine (UCI) Machine Learning Repository [34]. Table 1 displays the dataset information, including the specific numbers of training samples, test samples, and features. The best parameters of FS-NSVR, $\text{L}_1$-SVR, $\text{L}_\text{p}$-SVR, and $\text{L}_1$-LSSVR for real-world datasets are shown in Table 4.

Table 5 lists the feature selection and regression results of FS-NSVR, $\text{L}_1$-SVR, $\text{L}_\text{p}$-SVR, and $\text{L}_1$-LSSVR. One can easily observe that FS-NSVR selects fewer features than those of $\text{L}_1$-SVR, $\text{L}_\text{p}$-SVR, and $\text{L}_1$-LSSVR, but FS-NSVR obtains small values for NMSE and RMSE, which are comparable with the other methods. FS-NSVR selects very few useful features and captures the nonlinear statistical information in the test datasets. FS-NSVR realizes feature selection and regression simultaneously due to its inherent feature selection property. Faced with complex nonlinear high-dimensional datasets, $\text{L}_1$-SVR, $\text{L}_\text{p}$-SVR, and $\text{L}_1$-LSSVR present challenges since they can only solve the feature selection problem for the linear version. Regarding the running time, FS-NSVR is significantly faster than $\text{L}_1$-SVR and $\text{L}_\text{p}$-SVR.

Table 1: Specification of datasets.

| Data sets | No. of training samples | No. of testing samples | No. of features |
|---|---|---|---|
| Type A | 200 | 200 | 500 |
| Type B | 200 | 200 | 500 |
| Type C | 200 | 200 | 500 |
| Traffic | 1261 | 840 | 48 |
| Student | 106 | 289 | 27 |
| CNN | 500 | 618 | 81 |
| Community | 999 | 994 | 102 |
| Pakinson | 1040 | 612 | 26 |

Table 2: The optimal parameters of artificial datasets.

| Data sets | Regressor | $C$ | $\delta$ |
|---|---|---|---|
| Type A | FS-NSVR | $2^4$ | $2^3$ |
| | $L_1$-SVR | $2^{-5}$ | $2^0$ |
| | $L_p$-SVR | $2^{-3}$ | $2^{-3}$ |
| | $L_1$-LSSVR | $2^{-3}$ | $2^{-3}$ |
| Type B | FS-NSVR | $2^3$ | $2^2$ |
| | $L_1$-SVR | $2^0$ | $2^0$ |
| | $L_p$-SVR | $2^{-3}$ | $2^{-3}$ |
| | $L_1$-LSSVR | $2^{-1}$ | $2^{-1}$ |
| Type C | FS-NSVR | $2^1$ | $2^1$ |
| | $L_1$-SVR | $2^{-2}$ | $2^{-2}$ |
| | $L_p$-SVR | $2^{-4}$ | $2^{-1}$ |
| | $L_1$-LSSVR | $2^{-4}$ | $2^{-4}$ |

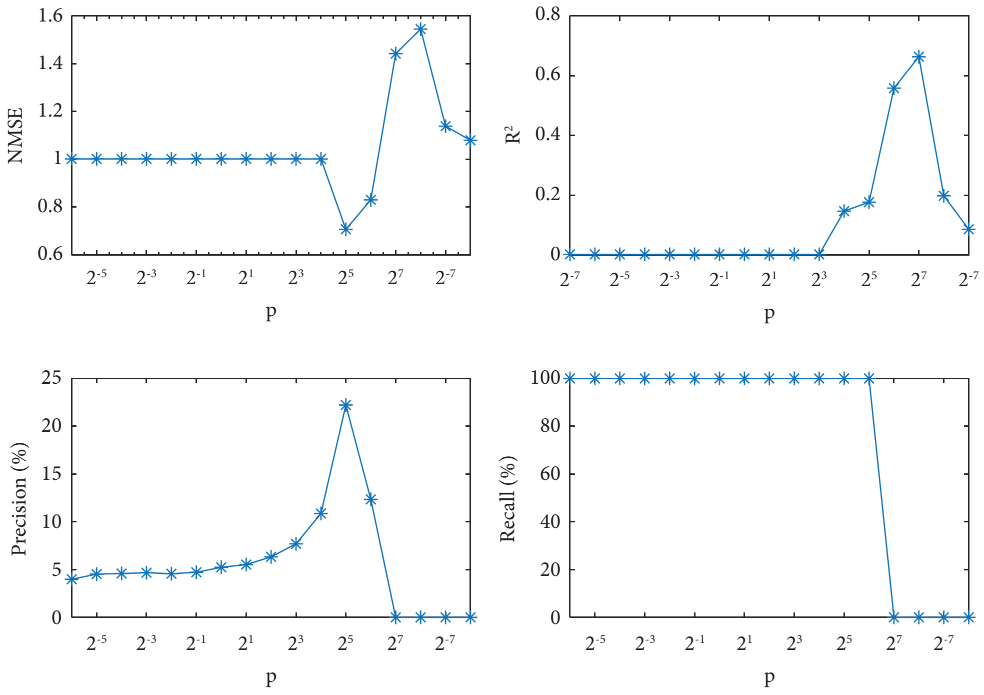

Figure 1: The influence of $p$ on NMSE, $R^2$, precision, and recall for the Type A dataset. Parameter $C$ is fixed as the optimal value.

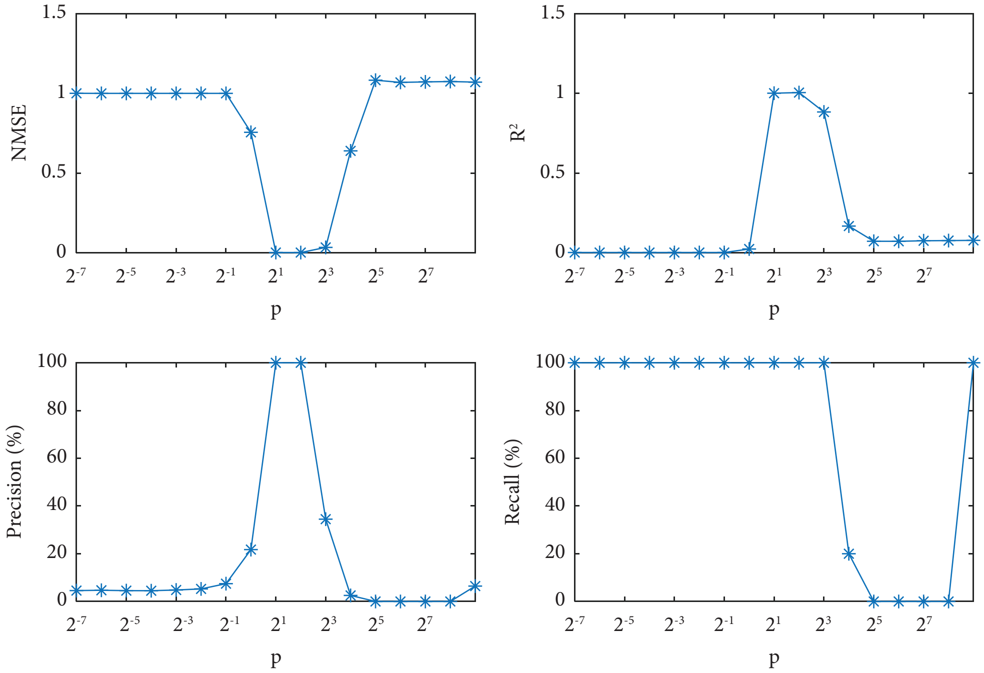

Figure 2: The influence of $p$ on NMSE, $R^2$, precision, and recall for the Type B dataset. Parameter $C$ is fixed as the optimal value.

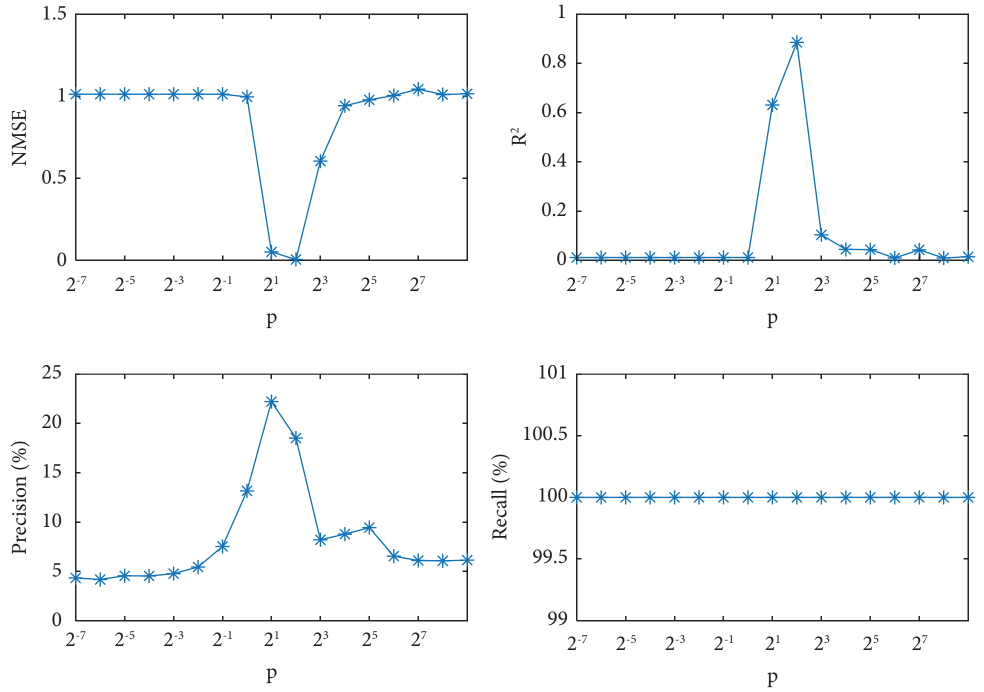

Figure 3: The influence of $p$ on NMSE, $R^2$, precision, and recall for the Type C dataset. Parameter $C$ is fixed as the optimal value.

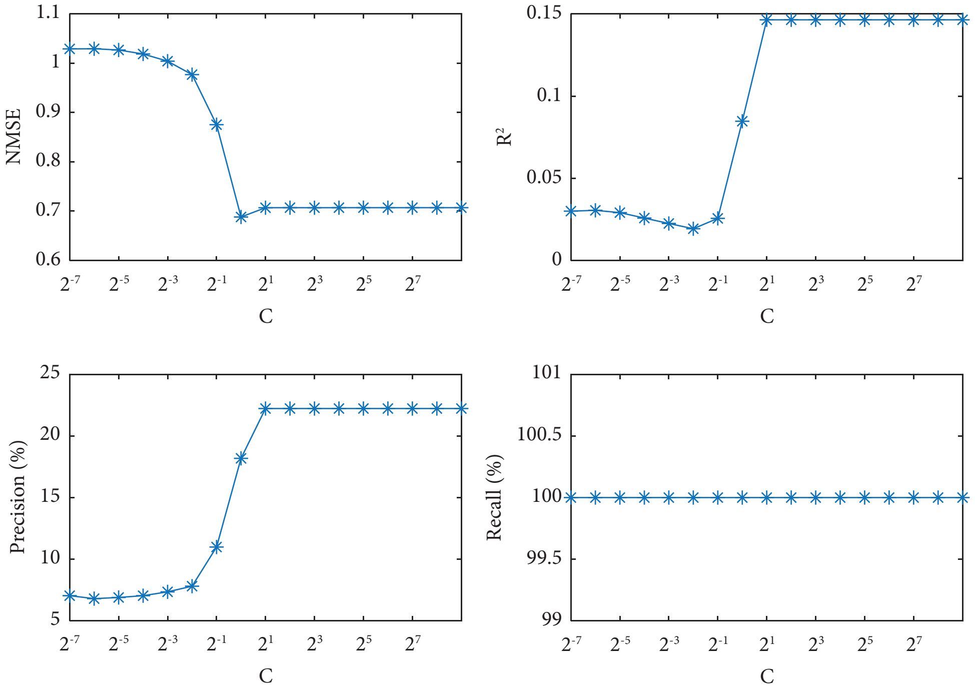

Figure 4: The influence of $C$ on NMSE, $R^2$, precision, and recall for the Type A dataset. Parameter $p$ is fixed as the optimal value.

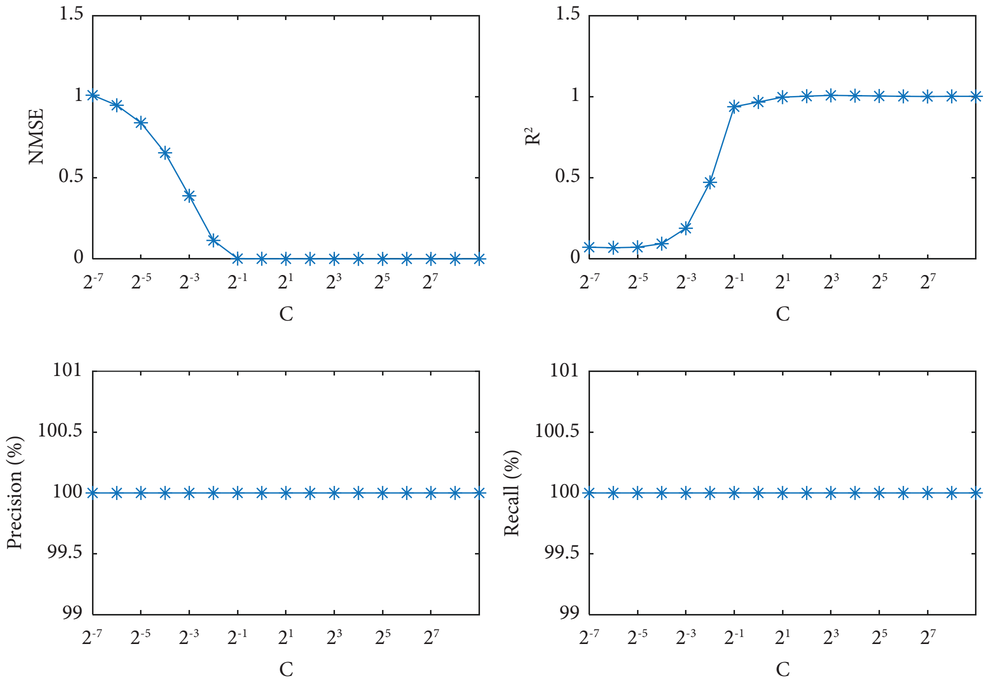

Figure 5: The influence of $C$ on the NMSE, $R^2$, precision, and recall for the Type B dataset. Parameter $p$ is fixed as the optimal value.

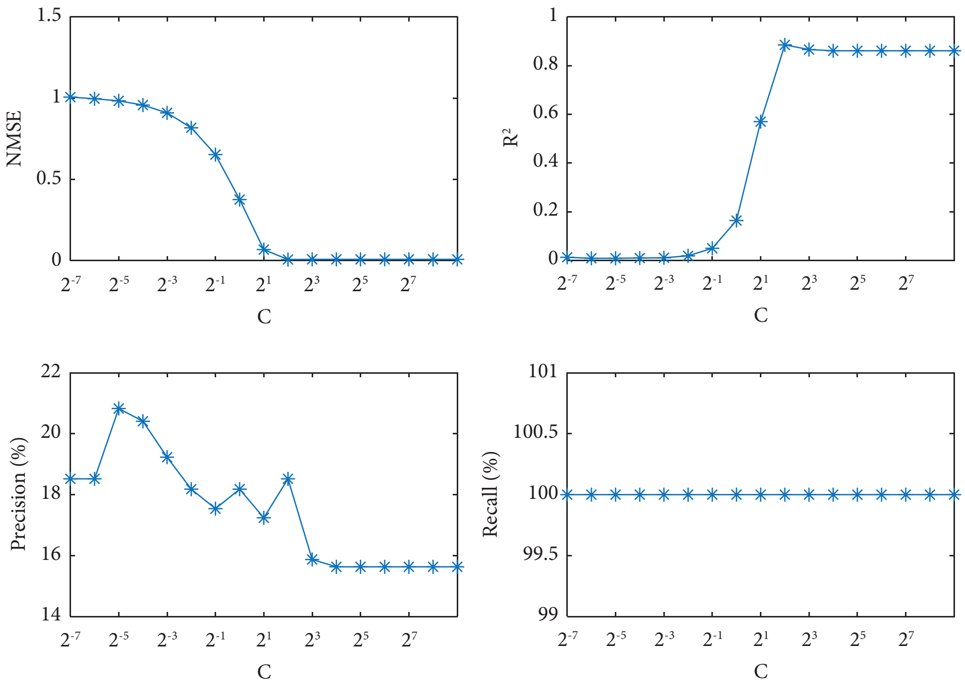

FIGURE 6: The influence of $C$ on the NMSE, $R^2$, precision, and recall for the Type C dataset. Parameter $p$ is fixed as the optimal value.

TABLE 3: Comparison results of FS-NSVR, $L_1$-SVR, $L_p$-SVR, and $L_1$-LSSVR for artificial datasets.

| Data sets | Regressor | NMSE | $R^2$ | RMSE | Precision | Recall | CPU S. |
|---|---|---|---|---|---|---|---|
| Type A | FS-NSVR | **0.707** | **0.146** | **0.146** | **22.22** | **100** | 0.148 |
| | $L_1$-SVR | 1.032 | 0.031 | 0.177 | 16.67 | 10 | 2.107 |
| | $L_p$-SVR | 1.028 | 0.028 | 0.176 | — | 0 | 2.701 |
| | $L_1$-LSSVR | 1.028 | 0.031 | 0.176 | 0 | 0 | **0.004** |
| Type B | FS-NSVR | **0.032** | **0.884** | **0.069** | **34.48** | **100** | 0.071 |
| | $L_1$-SVR | 1.068 | 0.074 | 0.396 | 1.26 | 10 | 1.379 |
| | $L_p$-SVR | 1.077 | 0.078 | 0.397 | — | 0 | 2.702 |
| | $L_1$-LSSVR | 1.001 | 0.001 | 0.383 | — | 0 | **0.004** |
| Type C | FS-NSVR | **0.006** | **0.885** | **0.324** | **18.51** | **100** | 0.076 |
| | $L_1$-SVR | 1.003 | 0.009 | 4.186 | 0 | 0 | 1.229 |
| | $L_p$-SVR | 1.011 | 0.011 | 4.201 | — | 0 | 2.584 |
| | $L_1$-LSSVR | 0.997 | 0.001 | 4.174 | 0 | 0 | **0.005** |

Bold values refer to the best performing regressors under each criterion.

TABLE 4: The optimal parameters of real-world datasets.

| Data sets | Regressor | $C$ | $\delta$ |
|---|---|---|---|
| Traffic | FS-NSVR | $2^{-1}$ | $2^{2}$ |
| | $L_1$-SVR | $2^{1}$ | $2^{1}$ |
| | $L_p$-SVR | $2^{-7}$ | $2^{8}$ |
| | $L_1$-LSSVR | $2^{4}$ | `$2^{4}$ |
| Student | FS-NSVR | $2^{7}$ | $2^{5}$ |
| | $L_1$-SVR | $2^{-1}$ | $2^{-1}$ |
| | $L_p$-SVR | $2^{-4}$ | $2^{4}$ |
| | $L_1$-LSSVR | $2^{-3}$ | $2^{-3}$ |
| CNN | FS-NSVR | $2^{8}$ | $2^{3}$ |
| | $L_1$-SVR | $2^{4}$ | $2^{4}$ |
| | $L_p$-SVR | $2^{-3}$ | $2^{1}$ |
| | $L_1$-LSSVR | $2^{5}$ | $2^{5}$ |

Table 4: Continued.

| Data sets | Regressor | $C$ | $\delta$ |
|---|---|---|---|
| Community | FS-NSVR | $2^5$ | $2^1$ |
| | $L_1$-SVR | $2^0$ | $2^0$ |
| | $L_p$-SVR | $2^{-1}$ | $2^0$ |
| | $L_1$-LSSVR | $2^4$ | $2^4$ |
| Pakinson | FS-NSVR | $2^5$ | $2^2$ |
| | $L_1$-SVR | $2^{-5}$ | $2^{-5}$ |
| | $L_p$-SVR | $2^{-3}$ | $2^3$ |
| | $L_1$-LSSVR | $2^6$ | $2^6$ |

Bold values refer to the best performing regressors under each criterion.

Table 5: Comparison results of FS-NSVR, $L_1$-SVR, $L_p$-SVR, and $L_1$-LSSVR for real-world datasets.

| Data sets | Regressor | $P_1$ (%) | NMSE | RMSE | $R^2$ | CPU S. |
|---|---|---|---|---|---|---|
| Student | FS-NSVR | **6.250** | 0.094 | 0.037 | 0.913 | 0.070 |
| | $L_1$-SVR | 72.917 | 0.062 | 0.030 | 0.956 | 1.118 |
| | $L_p$-SVR | 62.500 | 0.065 | 0.030 | **0.980** | 0.400 |
| | $L_1$-LSSVR | 52.083 | 0.062 | **0.025** | 0.960 | **0.001** |
| Traffic | FS-NSVR | **18.519** | **0.905** | **0.946** | 0.102 | 2.007 |
| | $L_1$-SVR | 85.185 | 0.963 | 0.976 | 0.252 | 2.259 |
| | $L_p$-SVR | 88.889 | 1.083 | 1.035 | **0.530** | 0.457 |
| | $L_1$-LSSVR | 44.444 | 0.952 | 1.602 | 0.057 | **0.001** |
| CNN | FS-NSVR | **7.407** | 0.001 | 0.035 | **1.000** | 0.237 |
| | $L_1$-SVR | 86.420 | 0.00 | **0.011** | **1.000** | 1.249 |
| | $L_p$-SVR | 9.877 | 0.001 | 0.032 | 0.981 | 0.372 |
| | $L_1$-LSSVR | 8.642 | 0.000 | 0.020 | 0.983 | **0.001** |
| Community | FS-NSVR | **21.569** | 0.379 | 0.138 | **0.918** | 1.108 |
| | $L_1$-SVR | 66.667 | 0.331 | **0.129** | 0.656 | 2.047 |
| | $L_p$-SVR | 37.255 | 0.337 | 0.130 | 0.637 | 1.822 |
| | $L_1$-LSSVR | 74.510 | 0.340 | 0.130 | 0.731 | **0.001** |
| Pakinson | FS-NSVR | **23.077** | **0.862** | 0.446 | 0.415 | 1.176 |
| | $L_1$-SVR | 42.308 | 1.093 | 0.502 | 0.319 | 1.276 |
| | $L_p$-SVR | 88.462 | 1.229 | 0.533 | **0.430** | 0.570 |
| | $L_1$-LSSVR | 76.923 | 1.021 | **0.372** | 0.197 | **0.001** |

## 5. Conclusion

Our paper focused on the nonlinear feature selection problem posed by high-dimensional data, especially when nonlinear complex relationships between features exist. To solve this problem, we brought a feature selection matrix in the original space into nonlinear support vector regression and then proposed a novel feature selection method for nonlinear support vector regression (FS-NSVR). FS-NSVR is a mixed-integer programming problem (MPP). To efficiently solve FS-NSVR, we employed an alternate iterative greedy algorithm to find a local optimal value. The feature selection matrix and the supervised selection process of FS-NSVR ensured that the representative features were selected and the redundant features were discarded automatically. The feature selection and regression performance of FS-NSVR in the artificial and real-world datasets confirmed its sparseness and effectiveness.

The proposed method also has limitations that should be acknowledged in future studies. First, FS-NSVR is not suitable for the heterogeneity problem of nonlinear high-dimensional data. The spirit of quantile regression [35–37] can be brought into the nonlinear feature selection framework in the future. Second, more efficient methods to solve FS-NSVR are needed since the training speed of the current method is not fast enough with regard to large-scale data sets. Third, forming the application perspective on how to use FS-NSVR to deal with nonlinear feature selection problem in the real world remains a question in future work.

## Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 11871183, 61866010, 61603338, and 12101552), the National Social Science Foundation of China (No. 21BJY256), the Philosophy and Social Sciences Leading Talent Training Project of Zhejiang Province (No. 21YJRC07-1YB), the Natural Science

Foundation of Zhejiang Province (No. LY21F030013), and the Natural Science Foundation of Hainan Province (No. 120RC449).

## References

[1] G. Kou, Y. Xu, Y. Peng et al., "Bankruptcy prediction for SMEs using transactional data and two-stage multiobjective feature selection," *Decision Support Systems*, vol. 140, Article ID 113429, 2021.

[2] S. Basu and G. Michailidis, "Regularized estimation in sparse high-dimensional time series models," *Annals of Statistics*, vol. 43, no. 4, pp. 1535–1567, 2015.

[3] M. Yamada, J. Tang, J. Lugo-Martinez et al., "Ultra high-dimensional nonlinear feature selection for big biological data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 7, pp. 1352–1365, 2018.

[4] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in Bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.

[5] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, "Detecting and quantifying causal associations in large nonlinear time series datasets," *Science Advances*, vol. 5, no. 11, Article ID eaau4996, 2019.

[6] G. Ling, A. Razzaq, Y. Guo, T. Fatima, and F. Shahzad, "Asymmetric and time-varying linkages between carbon emissions, globalization, natural resources and financial development in China," *Environment, Development and Sustainability*, vol. 24, no. 5, pp. 6702–6730, 2022.

[7] X. Zhao, W. Wang, L. Liu, and Y. C. T. Shih, "A flexible quantile regression model for medical costs with application to Medical Expenditure Panel Survey Study," *Statistics in Medicine*, vol. 37, no. 17, pp. 2645–2666, 2018.

[8] K. Wang, Q. Zhao, J. Lu, and T. Yu, "Profiles: A Nonlinear Clustering Method for Pattern Detection in High Dimensional Data," *BioMed research international*, vol. 2015, Article ID 918954, 2015.

[9] A. Tayal, T. F. Coleman, and Y. Li, "Primal explicit max margin feature selection for nonlinear support vector machines," *Pattern Recognition*, vol. 47, no. 6, pp. 2153–2164, 2014.

[10] X. Yan, Y. Liu, P. Ding, and M. Jia, "Fault Diagnosis of Rolling-Element Bearing Using Multiscale Pattern Gradient Spectrum Entropy Coupled with Laplacian Score," *Complexity*, vol. 2020, Article ID 4032628, , 2020.

[11] X. Yan, Y. Liu, and M. Jia, "A feature selection framework-based multiscale morphological analysis algorithm for fault diagnosis of rolling element bearing," *IEEE Access*, vol. 7, pp. 123436–123452, 2019.

[12] M. Luo, X. Chang, L. Nie, Y. Yang, A. G. Hauptmann, and Q. Zheng, "An adaptive semisupervised feature analysis for video semantic recognition," *IEEE Transactions on Cybernetics*, vol. 48, no. 2, pp. 648–660, 2018.

[13] C. Yan, X. Chang, M. Luo et al., "Self-weighted robust LDA for multiclass classification with edge classes," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 12, no. 1, pp. 1–19, 2021.

[14] En Yu, J. Sun, J. Li, X. Chang, X. H. Han, and A. G. Hauptmann, "Adaptive semi-supervised feature selection for cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 21, no. 5, pp. 1276–1288, 2019.

[15] Zi-J. Wang, Z. H. Zhan, Y. Lin et al., "Automatic niching differential evolution with contour prediction approach for multimodal optimization problems," *IEEE Transactions on Evolutionary Computation*, vol. 24, no. 1, pp. 114–128, 2020.

[16] Zi-J. Wang, Z. H. Zhan, W. J. Yu et al., "Dynamic group learning distributed particle swarm optimization for large-scale optimization and its application in cloud workflow scheduling," *IEEE Transactions on Cybernetics*, vol. 50, no. 6, pp. 2715–2729, 2020.

[17] Zi-J. Wang, Z. H. Zhan, S. Kwong, H. Jin, and J. Zhang, "Adaptive granularity learning distributed particle swarm optimization for large-scale optimization," *IEEE Transactions on Cybernetics*, vol. 51, no. 3, pp. 1175–1188, 2021.

[18] Y. F. Ye, Y. H. Shao, N. Y. Deng, C. N. Li, and X. Y. Hua, "Robust L-norm least squares support vector regression with feature selection," *Applied Mathematics and Computation*, vol. 305, pp. 32–52, 2017.

[19] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 245–271, 1997.

[20] J. Bi, K. Bennett, M. Embrechts, C. Breneman, and M. Song, "Dimensionality reduction via sparse support vector machines," *Journal of Machine Learning Research*, vol. 3, no. Mar, pp. 1229–1243, 2003.

[21] Z. Y. Algamal, M. K. Qasim, and H. T. M. Ali, "A QSAR classification model for neuraminidase inhibitors of influenza a viruses (H1N1) based on weighted penalized support vector machine," *SAR and QSAR in Environmental Research*, vol. 28, no. 5, pp. 415–426, 2017.

[22] M. K. Qasim, Z. Y. Algamal, and H. M. Ali, "A binary QSAR model for classifying neuraminidase inhibitors of influenza A viruses (H1N1) using the combined minimum redundancy maximum relevancy criterion with the sparse support vector machine," *SAR and QSAR in Environmental Research*, vol. 29, no. 7, pp. 517–527, 2018.

[23] O. S. Qasim and Z. Y. Algamal, "A gray wolf algorithm for feature and parameter selection of support vector classification," *International Journal of Computing Science and Mathematics*, vol. 13, no. 1, pp. 93–102, 2021.

[24] Z. Y. Algamal, M. K. Qasim, M. H. Lee, and H. T. M. Ali, "Improving grasshopper optimization algorithm for hyperparameters estimation and feature selection in support vector regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 208, Article ID 104196, 2021.

[25] O. M. Ismael, O. S. Qasim, and Z. Y. Algamal, "Improving Harris hawks optimization algorithm for hyperparameters estimation and feature selection in v-support vector regression based on opposition-based learning," *Journal of Chemometrics*, vol. 34, no. 11, Article ID e3311, 2020.

[26] O. M. Ismael, O. S. Qasim, and Z. Y. Algamal, "A new adaptive algorithm for v-support vector regression with feature selection using Harris hawks optimization algorithm," *Journal of Physics: Conference Series*, vol. 1897, Article ID 012057, 2021.

[27] X. Peng and D. Xu, "A local information-based feature-selection algorithm for data regression," *Pattern Recognition*, vol. 46, no. 9, pp. 2519–2530, 2013.

[28] Y. F. Ye, C. Ying, Y. X. Jiang, and C. N. Li, "L1-norm least squares support vector regression via the alternating direction method of multipliers," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 21, no. 6, pp. 1017–1025, 2017.

[29] C. Zhang, D. Li, and J. Tan, "The support vector regression with adaptive norms," *Procedia Computer Science*, vol. 18, pp. 1730–1736, 2013.

[30] Y. A. Guzman, *Theoretical Advances in Robust Optimization, Feature Selection, and Biomarker Discovery*, Doctoral

dissertation Princeton University, Princeton, NJ 08544, USA, 2016.

[31] M. Onel, C. A. Kieslich, Y. A. Guzman, C. A. Floudas, and E. N. Pistikopoulos, “Reprint of: big data approach to batch process monitoring: simultaneous fault detection and diagnosis using nonlinear support vector machine-based feature selection,” *Computers & Chemical Engineering*, vol. 116, pp. 503–520, 2018.

[32] O. L. Mangasarian and G. Kou, “Feature Selection for Nonlinear Kernel Support Vector Machines,” in *Proceedings of the Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007)*, pp. 231–236, Omaha, NE, USA, March 2007.

[33] B. Lan, Z. Wang, Y.-H. Shao, and N.-Y. Deng, “A Novel Feature Selection Method for Twin Support Vector Machine,” *Knowledge-Based Systems*, vol. 59, 2014.

[34] C. Blake, “UCI Repository of Machine Learning Databases,” *Machine Learning Repository*, 1998.

[35] R. Koenker and G. Bassett, “Regression Quantiles,” *journal of the Econometric Society*, vol. 1, pp. 33–50, 1978.

[36] Y. Ye, Y. Shao, and C. Li, “A sparse approach for high-dimensional data with heavy-tailed noise,” *Economic Research-Ekonomska Istraživanja*, vol. 35, no. 1, pp. 2764–2780, 2022.

[37] Y. Ye, Y. Shao, C. Li, X. Hua, and Y. Guo, “Online support vector quantile regression for the dynamic time series with heavy-tailed noise,” *Applied Soft Computing*, vol. 110, Article ID 107560, 2021.