A Modified Robust Support Vector Regression Approach for Data Containing High Leverage Points and Outliers in the Y-direction

Habshah Midi¹, Jama Mohamed^{2,*}

¹Faculty of Science and Institute for Mathematical Research, University Putra Malaysia, Malaysia

²Faculty of Mathematics and Statistics, College of Applied and Natural Science, University of Hargeisa, Somaliland

Received May 24, 2020; Revised July 3, 2020; Accepted July 29, 2020

Cite This Paper in the following Citation Styles

(a): [1] Habshah Midi, Jama Mohamed, "A Modified Robust Support Vector Regression Approach for Data Containing High Leverage Points and Outliers in the Y-direction," *Mathematics and Statistics*, Vol. 8, No. 5, pp. 493 - 505, 2020. DOI: 10.13189/ms.2020.080502.

(b): Habshah Midi, Jama Mohamed (2020). A Modified Robust Support Vector Regression Approach for Data Containing High Leverage Points and Outliers in the Y-direction. *Mathematics and Statistics*, 8(5), 493 - 505. DOI: 10.13189/ms.2020.080502.

Copyright©2020 by authors, all rights reserved. Authors agree that this article remains permanently open access under the terms of the Creative Commons Attribution License 4.0 International License

Abstract The support vector regression (SVR) model is currently a very popular non-parametric method used for estimating linear and non-linear relationships between response and predictor variables. However, there is a possibility of selecting vertical outliers as support vectors that can unduly affect the estimates of regression. Outliers from abnormal data points may result in bad predictions. In addition, when both vertical outliers and high leverage points are present in the data, the problem is further complicated. In this paper, we introduced a modified robust SVR technique in the simultaneous presence of these two problems. Three types of SVR models, i.e. eps-regression (ϵ -SVR), nu-regression (ν -SVR) and bound constraint eps-regression (ϵ -BSVR), with eight different kernel functions are integrated into the new proposed algorithm. Based on 10-fold cross-validation and some model performance measures, the best model with a suitable kernel function is selected. To make the selected model robust, we developed a new double SVR (DSVR) technique based on fixed parameters. This can be used to detect and reduce the weight of influential observations or anomalous points in the data set. The effectiveness of the proposed technique is verified by using a simulation study and some well-known contaminated data sets.

Keywords Double Support Vector Regression, Fixed Parameters, Vertical Outliers, High Leverage Points, Robust Mahalanobis Distance

1. Introduction

The SVR model is first introduced by [1]. It depends on the powerful concept of support vector machine (SVM), which is a typical perspective in the field of statistical learning. With a specific and global optimal solution, SVR has become popular in recent years due to its excellent performance and high generalization capability [2]. Some of the reasons for the widespread use of the SVR include reduced sensitivity to relative minima, hypothetical output guarantees, and a high degree of versatility to add additional dimensions to the input space, thereby avoiding the model's increasing complexity [3].

The superiority of SVR is its ability to approximate nonlinear relationships by the use of kernel tricks by constructing a sparse model to tackle regression problems [3]. To explain this, consider a data set such that an original input is $x \in R^n$ and a target output is $y \in R^1$. The input vector x is first mapped onto a high-dimensional feature space in the SVR method, which is nonlinearly connected to the input space. The idea is to use the kernel trick in a high-dimensional predictor space to approximate the nonlinear relationship within the original input space to linear form [4–5]. The function of regression is given by:

$$f(x) = \langle w, \Phi(x) \rangle + b \quad (1)$$

where $\Phi(x)$ is a non-linear function, w and b correspond to the weight and the bias. The goal of SVR is to estimate the values of parameters w and b , which optimize the expected risk by minimizing the ε -insensitive loss function below:

$$L_\varepsilon(y_i) = \begin{cases} 0, & \text{if } |y_i - w^T x_i| \leq \varepsilon \\ |y_i - w^T x_i| - \varepsilon, & \text{otherwise} \end{cases} \quad (2)$$

In other words, SVR tries to minimize the bound generalization error such that generalized efficiency is achieved instead of minimizing the actual training error [6]. It is described through the use of kernel functions, sparse solution, and margin and number of support vectors controlled by Vapnik-Chervonenkis (VC) theory [7]. The key terminologies in SVR are shown in Figure 1.

Figure 1. Basic terminology of SVR

According to [8], most applications for real-world data are subject to anomalies and noises, which is a common issue that leads to misleading and false conclusions. Barnett and Lewis [9] have described an outlier as one that appears to significantly deviate from other members of the sample it occurs in. Outliers are observations which are far from the bulk of the rest of the data [10]. In SVR, it is possible to select outlying observations (outliers and high leverage points) as supporting vectors which may affect the estimation process [11–12]. It is, therefore, useful to implement some robust methods in SVR to remedy both problems of outliers and high leverage points.

Several studies were conducted by utilizing SVR to analyze data consisting of outliers. For instance, Jordaan and Smits [13], using the benefits of the Lagrange multipliers, developed the traditional SVR method for outlier detection in relation to Karush-Kuhn-Tucker (KKT) requirements. The main drawbacks of this approach are high calculation costs, difficult operation for non-expert users and the possibility of the emergence of masking and swamping problems. To solve the problems of the standard

method, [14] suggested the μ - ε -SVR outlier detection procedure, which uses a new parameter of regularization (abbreviated as μ). The possible shortcomings of this method are detection of one outlier per iteration, high computational costs and lack of clear rule for selecting for the threshold value, ε which complicates the approach. In order to avoid the disadvantages of the above techniques, [15] presented a functional technique for the detection of outliers, using non-sparse ε -SVR that helps to minimize the time cost and incorporates fixed parameters. The robust SVR based on this approach is often referred to as Double SVR (DSVR). The objective of this study is to modify the DSVR technique by integrating three types of SVR models, i.e. epsilon-SVR (ε -SVR), nu-SVR (ν -SVR), and epsilon bound-constrained SVR (ε -BSVR), with different kernel functions into the algorithm.

2. Methodology

2.1. Standard SVR Models

The SVR concept is suggested by [1] to solve problems of function fitting, which is called the ε -SVR. Given a set of data points, $\{(x_1, y_1), \dots, (x_n, y_n)\}$ such that $x_i \in R^n$ is an input and $y_i \in R^1$ is a target output, we want to employ ε -SVR. The optimization problem is:

$$\begin{aligned} \text{Minimize} \quad & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) \\ \text{subject to} \quad & \begin{cases} y_i - w^T x_i - b \leq \varepsilon + \xi_i \\ w^T x_i + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0, i = 1, 2, \dots, n \end{cases} \end{aligned} \quad (3)$$

This can be solved as a quadratic programming problem. C is the parameter to control the amount of influence of the error, ξ_i, ξ_i^* are slack variables and b is the bias term.

In [16], the authors introduced an improved SVR model, called nu-SVR (ν -SVR), which uses a ' ν ' parameter to monitor the amount of training error and support vectors. To put it in another way, the parameter ν is used to evaluate the proportion of the number of support vectors we want to hold in our solution about the total number of samples in the dataset. The size of ν is between 0 and 1. That is, $0 \leq \nu \leq 1$. After substituting ε for ν , (3) is modified as can be seen in (4).

$$\begin{aligned} \text{Minimize} \quad & \frac{1}{2} \|w\|^2 + C \left(\nu \varepsilon + \frac{1}{n} \sum_{i=1}^n (\xi_i + \xi_i^*) \right) \\ \text{subject to} \quad & \begin{cases} y_i - w^T x_i - b \leq \varepsilon + \xi_i \\ w^T x_i + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0, i = 1, 2, \dots, n \end{cases} \end{aligned} \quad (4)$$

The difference between ε -SVR and ν -SVR is that in

ν -SVR, we can control the number of support vectors, but in ε -SVR, we can control the amount of error in the model instead of support vectors.

Finally, the proposed method in [17], in which the bias term squared is applied to the objective function, is known as epsilon bound-constrained SVR (ε -BSVR) formulation. The optimization problem of ε -BSVR is:

$$\begin{aligned} &\text{Minimize} \quad \frac{1}{2} \|w\|^2 + \frac{1}{2} b^2 + \frac{C}{n} \sum_{i=1}^n (\xi_i + \xi_i^*) \\ &\text{subject to} \quad \begin{cases} y_i - w^T x_i - b \leq 1 + \xi_i \\ w^T x_i + b - y_i \leq 1 + \xi_i^* \\ \xi_i, \xi_i^* \geq 0, i = 1, 2, \dots, n \end{cases} \end{aligned} \quad (5)$$

This can be solved with Wolfe's dual to a quadratic programming problem with box constraints only. The parameters of ε -BSVR are the same as the parameters of ε -SVR except that the parameter b , the bias term, is

restricted to a given range.

2.2. Kernel Functions

To employ SVR models, kernel functions are to be used. The kernel function measures the dot product of two vectors x_i and x_j in the feature space $\Phi(x_i)$ and $\Phi(x_j)$. That is,

$$K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j) \quad (6)$$

In SVR procedure, $K(x_i, x_j)$ is used in some form of nonlinear relationship to map the input space into high dimensional predictor space. Several kernel function types are employed in SVR. The type of kernel function chosen affects the estimates and computational complexity for implementing the designed SVR. Table 1 outlines eight of the common kernel function types that are considered for this research study.

Table 1. Eight kernel functions considered for the study

Kernel type	Function
Linear	$k(x_i, x) = \langle x_i, x \rangle$
RBF or Gaussian	$k(x_i, x) = \exp(-\sigma \ x_i - x\ ^2)$
Polynomial	$k(x_i, x) = (\text{scale} \cdot \langle x_i, x \rangle + \text{offset})^{\text{degree}}$
Hyperbolic Tangent	$k(x_i, x) = \tanh(\text{scale} \cdot \langle x_i, x \rangle + \text{offset})$
Bessel function	$k(x_i, x) = \frac{Bessel_{(v+1)}^n(\sigma \ x_i - x\)}{(\ x_i - x\)^{-n(v+1)}}$
Laplace RBF	$k(x_i, x) = \exp(-\sigma \ x_i - x\)$
ANOVA Radial Basis	$k(x_i, x) = \left(\sum_{k=1}^n \exp(-\sigma(x_i^k - x^k)^2) \right)^d$
Spline	<p>One dimension:</p> $k(x_i, x) = 1 + x_i x \min(x_i, x) - \frac{x_i + x}{2} (\min(x_i, x)^2 + \frac{\min(x_i, x)^3}{3})$ <p>multidimensional:</p> $k(x_i, x) = \prod_{k=1}^n k(x^k, x^k)$

2.3. Double SVR (DSVR) Algorithm

The Double SVR (DSVR) is proposed by [12]. It is a practical method for detecting influential observations by considering three directions: type of transformation, sparseness and robustness. The efficiency of this approach arises from the fact that it decreases the computational cost and can identify outliers without having to delete them to enable handling. This can be achieved by minimizing the weights for outlying points in the data set. The DSVR is based on fixed parameters SVR (FPSVR) method [8]. The benefit of this technique is to control the free parameters, ε , C and σ , where σ is the hyperparameter of the RBF kernel.

The method of DSVR based on fixed parameters can be summarized as follows:

Step 1: Based on FPSVR (the fixed parameters of ε -SVR are: $\varepsilon = 0$, $C = 10000$ and $\sigma = 1$), find the fitted values of the ε -SVR model.

$$z = y_{\varepsilon-SVR} = \sum_{i=1}^n (\alpha_i - \alpha_i^*) k(x_i, x) + b \quad (7)$$

where $\alpha_i, \alpha_i^* \in [0, C]$ are Lagrange Multipliers, $k(x_i, x)$ is the kernel function and b is the constant (bias term). Or alternatively, find the absolute residuals based on the FPSVR (the fixed parameters are $\varepsilon = 0$, $C = 0.0001$ and $\sigma = 1$ in the case of using residuals, z). The absolute residuals are given in (8).

$$z = |y - y_{\varepsilon-SVR}| \quad (8)$$

Step 2: Any point with an absolute z value larger than the cutoff point is considered to be an outlier. The cutoff point is given in (9).

$$CP = 2\text{median}(z) + 2\sqrt{\frac{\pi \text{var}(z)}{2n}} \quad (9)$$

Step 3: Compute the weight function as given in (10).

$$\tau_i = \min[1, CP / z] \quad (10)$$

Step 4: Estimate the final robust ε -SVR by substituting (10) into (7) as follows:

$$f(x) = \sum_{i=1}^n \tau_i (\alpha_i - \alpha_i^*) k(x_i, x) + b \quad (11)$$

2.4. The Proposed Robust SVR Method

As the standard SVR models are not good robust against outliers and high leverage points, DSVR is developed. Nevertheless, the DSVR considers only ε -SVR with RBF kernel. Hence, we improved the DSVR by combining all three forms of SVR into the algorithm with different kernel functions. The hyperparameter of any kernel is abbreviated as h . In this method, the value of h is set equal to 1. If the kernel function has more than one hyperparameter, like the

polynomial kernel, they all need to be equal to 1. The proposed robust SVR approach based on fixed parameters can be summarized as follows:

Step 1: Based on FPSVR (in ε -SVR and ε -BSVR the fixed parameters are $\varepsilon = 0$, $C = 10,000$ and $h = 1$, and in ν -SVR, $\varepsilon = 0$, $\nu = 1$ and $h = 1$), find the fitted values of the selected SVR model.

$$z = y_{SVR} = \sum_{i=1}^n (\alpha_i - \alpha_i^*) k(x_i, x) + b \quad (12)$$

Or alternatively, based on the FPSVR (in ε -SVR and ε -BSVR the fixed parameters are $\varepsilon = 0$, $C = 0.0001$ and $h = 1$, and in ν -SVR, $\varepsilon = 0$, $\nu = 0.01$ and $h = 1$), find the absolute residuals.

$$z = |y - y_{SVR}| \quad (13)$$

The values of the parameters C and ν change with (12) and (13) because in both equations we considered that the fitted and estimated error values to be large enough to detect the vertical outliers.

Step 2: Any point with an absolute z value larger than the cutoff point is considered to be an outlier.

$$CP = \text{Median}(z) + 3\text{MAD}(z) \quad (14)$$

where $\text{MAD}(z) = b\text{Med}\{|z_i - \text{Med}(z)|\}$ and b is 1.4826 for normal distribution. For other distributions, b is the inverse of 75th percentile (Q_3) of the raw MAD [18–19].

Step 3: The suspect high leverage points are detected by using robust Mahalanobis distance (RMD) based on the Minimum Volume Ellipsoid (MVE) developed by [20] as:

$$\text{RMD}_i = \sqrt{(X - T_R(X))^T C_R(X)^{-1} (X - T_R(X))}, \text{ for } i = 1, 2, \dots, n \quad (15)$$

where $T_R(X)$ and $C_R(X)$ are robust locations and shape estimates of the MVE, respectively. Rahmatullah Imon [21] suggested a cut-off point for the robust Mahalanobis distances as:

$$CP = \text{Median}(\text{RMD}_i) + 3\text{MAD}(\text{RMD}_i) \quad (16)$$

The RMD is only valid for small dimensional data ($P < n$), where P is the number of predictor variables and n is the number of data points.

Step 4: Compute the weight function:

$$w_i = \begin{cases} \min[1, CP / z] & \text{for outlier cases} \\ \min[1, CP / \text{RMD}_i] & \text{for HLPs} \end{cases} \quad (17)$$

where z is the estimated, predicted or fitted values, as in (12), and HLPs are high leverage points.

Step 5: Estimate the final robust SVR model as follows:

$$f(x) = \sum_{i=1}^n w_i (\alpha_i - \alpha_i^*) k(x_i, x) + b \quad (18)$$

2.5. Performance Measures

To measure the accuracy and determine the best SVR model, this study considers the following three types of errors as stated in [22]:

1. Training error is the error that we get back to training data when we run the trained model.
2. Test error is the error that we get on the test data when we run the trained model.
3. A k-fold cross-validation error is an average error that we get when we run the trained model on k subsets (folds) of the test data.

Also, we can determine the goodness of the model by using the following measures:

1. The R-square (R^2) demonstrates the percentage of the total variation of the response variable that is explained by the predictor variable/s.

$$R^2 = \left(1 - \frac{RSS}{TSS}\right) \times 100 \quad (19)$$

where RSS is Residual Sum of Squares and TSS is Total Sum of Squares.

2. The Predicted R-squared (R^2_{pred}) reveals how well a regression model predicts outcomes for new data.

$$R^2_{pred} = \left(1 - \frac{PRESS}{TSS}\right) \times 100 \quad (20)$$

where $PRESS$ is Predicted Residual Error Sum of Squares.

$$PRESS = \sum_{i=1}^n (y_i - y_{i,-i})^2 \quad (21)$$

Finally, the assessment of the merits of the selected SVR, the DSVR and the proposed robust SVR methods can be made by using the following performance measures:

1. Mean Square Error (MSE) is used as an efficiency test to figure out which methods are the best in various cases.

$$MSE = \frac{\sum_{i=1}^n (y_i - y)^2}{n} \quad (22)$$

where y is the fitted values and n is the number of data points.

2. Root Mean Square Error (RMSE) indicates the closeness of the data points to the predicted values of the model. Lower RMSE values indicate improved fit.

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - y)^2}{n}} \quad (23)$$

3. Results and Discussion

3.1. Belgian Phone Data Set

Belgian phone data, collected from the Belgian Statistical Survey, is a real data set that has vertical outliers. It consists of the total number of international phone calls made between the years 1950 and 1973 (in ten million calls) [20]. The cross-validation error is determined using 10-fold cross-validation. We only used it to figure out the best model for this data set.

The results in Tables 2, 3 and 4 present the performance of the three different SVR types with eight different kernel functions of Belgian phone dataset by tuning the parameters ε , C and ν with 10-fold cross-validation in the grid search. The optimum parameters obtained in ε -SVR and ε -BSVR by the grid-search are $\varepsilon = 0.3$ and $C = 251$, while in ν -SVR, the optimum parameters are $\varepsilon = 0.4$ and $\nu = 0.1$. In these results, we can find that when Laplace kernel is used, the MSE is at least. In addition, the determination coefficient (R^2) is presented to measure the performance (goodness) of each model. In this data set, ε -SVR with Laplace kernel is the best model since MSE and cross-validation errors are minimal. As for the other SVR models, R^2 is also relatively high.

Figure 2 displays the outlier detection of ε -SVR technique with Laplace kernel depending on fixed parameters ($C = 10,000$, $\sigma = 1$, and $\varepsilon = 0$) by using the median absolute deviation of the absolute fitted values. We can see that the cases 15-20 are vertical outliers.

In Figure 3, we compared the performance of the three methods by considering different values for the parameters ($\varepsilon = 0.1, 0.2, 0.3$, $C = 1, 20, 50$ and $\sigma = 0.5, 1, 5$). The values of C greater than 50 give approximately the same results like those used. Furthermore, Figure 3 shows that ε -SVR is less efficient in reducing the estimation effects of the outliers, whereas the proposed robust SVR and DSVR methods compete with each other. However, in this data set, our proposed method is marginally better than DSVR by attaining the lowest MSE and RMSE values.

Table 2. ϵ -SVR ($\epsilon = 0.3$, $C = 251$) of Belgian phone data set

Kernel Type	Linear	Poly	RBF	Tanh	Laplace	Bessel	ANOVA	Spline
No. of Support Vectors Used	13	13	13	24	24	16	9	19
MSE	36.860	36.860	5.307	1.1×10^7	3.8261	16.562	5.861	1.2×10^5
Cross Validation Error	39.118	39.529	20.561	1.2×10^7	12.882	48.754	9.726	5.2×10^4
R^2	0.296	0.296	0.8711	0.004	0.990	0.608	0.860	0.062

Table 3. ν -SVR ($\epsilon = 0.4$, $\nu = 0.1$) of Belgian phone data set

Kernel Type	Linear	Poly	RBF	Tanh	Laplace	Bessel	ANOVA	Spline
No. of Support Vectors Used	4	4	10	4	18	4	5	4
MSE	56.119	56.119	30.501	42.775	34.664	55.622	45.006	49.173
Cross Validation Error	55.493	56.157	33.513	53.721	38.373	56.366	47.303	48.729
R^2	0.296	0.296	0.846	0.292	0.845	0.343	0.737	0.295

Table 4. ϵ -BSVR ($\epsilon = 0.3$, $C = 251$) of Belgian phone data set

Kernel Type	Linear	Poly	RBF	Tanh	Laplace	Bessel	ANOVA	Spline
No. of Support Vectors Used	13	13	14	24	24	16	9	20
MSE	36.864	36.864	3.4204	1.3×10^7	3.862	16.589	5.942	4.2×10^5
Cross Validation Error	46.320	46.731	17.504	1.5×10^7	41.738	26.773	9.356	2.1×10^5
R^2	0.296	0.296	0.937	0.006	0.989	0.6077	0.859	0.060

Figure 2. Outlier detection of ϵ -SVR with Laplace kernel of Belgian phone data set

Figure 3. The MSE and RMSE of ε -SVR, DSVR and the proposed robust SVR methods for Belgian phone data set

3.2. Hawkins-Bradu-Kass Data Set

In this example, we used the well-known artificial data set created by [23]. It consists of 75 observations with one dependent and three independent variables. Additionally, the first 14 cases are influential observations. 70 percent of this data set is randomly selected to use in training data and the remaining 30 percent in test data. Also, the cross-validation error is determined using 10-fold cross-validation. We only used this to determine the appropriate model and we used the complete data set for the rest.

Tables 5, 6 and 7 show the results of three different forms of SVR with eight different kernel functions of HBK data set by tuning the parameters ε , C and ν in the grid search with a 10-fold cross-validation error. The optimum parameters obtained in ε -SVR and ε -BSVR are $\varepsilon = 0$ and $C = 15$, while the optimum parameters in ν -SVR are $\varepsilon = 0.2$ and $\nu = 0.1$. Three types of errors, i.e. training, test and cross-validation errors, are considered to optimize the accuracy of the selection of the model. From the results, we can see that when using RBF or Gaussian

kernel, those errors are small. Moreover, Predicted R^2 (R^2 of the test data) is provided to determine the goodness of fit of each model. Based on these findings, the suitable SVR model of this data set is ε -BSVR with RBF kernel as the test and cross-validation errors are minimal and predicted R^2 is reasonably large when compared to the other SVR models.

The results in Figure 4, using RMD based on MVE, explain the detection of high leverage points for HBK data set. We can observe that the cases 1-14 are detected as influential observations.

In Figure 5, we considered different values for the parameters ($\varepsilon = 0.1, 0.2, 0.3$, $C = 1, 50, 100$ and $\sigma = 0.5, 1, 5$) for evaluating the three methods' performances. The values of C greater than 100 yield approximately the same results as those previously. Furthermore, the results in Figure 5 clearly shows the efficiency of the proposed robust SVR approach over DSVR and ε -BSVR techniques in terms of achieving lower MSE and RMSE values. Based on these outcomes, we can infer that the use of the proposed robust SVR method is recommended for HBK data set.

Table 5. ϵ -SVR ($\epsilon = 0$, $C = 15$) of HBK data set

Kernel Type	Linear	Poly	RBF	Tanh	Laplace	Bessel	ANOVA	Spline
No. of Support Vectors Used	53	53	53	53	53	53	53	53
MSE (Training)	6.988	6.987	0.025	2.3×10^3	0.000	0.527	0.333	0.299
MSE(Test)	5.800	5.800	0.974	1.5×10^3	7.990	1.360	0.853	1.086
Cross Validation Error	9.426	8.761	1.488	3.1×10^3	7.534	3.053	0.940	4.213
R^2_{pred}	0.670	0.670	0.931	0.181	0.731	0.913	0.948	0.931

Table 6. ν -SVR ($\epsilon = 0.2$, $\nu = 0.1$) of HBK data set

Kernel Type	Linear	Poly	RBF	Tanh	Laplace	Bessel	ANOVA	Spline
No. of Support Vectors Used	8	8	18	7	28	9	10	12
MSE (Training)	15.325	16.325	4.120	8.785	14.33	9.701	0.548	0.350
MSE (Test)	16.260	16.260	5.183	10.382	18.563	10.785	1.187	1.150
Cross Validation Error	15.786	15.222	6.0804	9.641	17.520	10.803	1.267	0.621
R^2_{pred}	0.672	0.673	0.946	0.419	0.847	0.821	0.926	0.928

Table 7. ϵ -BSVR ($\epsilon = 0$, $C = 15$) of HBK data set

Kernel Type	Linear	Poly	RBF	Tanh	Laplace	Bessel	ANOVA	Spline
No. of Support Vectors Used	53	53	53	53	53	53	53	53
MSE (Training)	6.499	6.491	0.082	2.0×10^3	0.000	1.056	0.254	0.299
MSE (Test)	9.443	9.508	0.636	2.1×10^3	5.550	0.892	0.797	1.086
Cross Validation Error	12.235	11.100	1.190	1.7×10^3	6.733	2.368	1.237	1.086
R^2_{pred}	0.654	0.654	0.956	0.001	0.827	0.927	0.937	0.931

Figure 4. Detection of high leverage points based on robust Mahalanobis distance for HBK data set

Figure 5. The MSE and RMSE of ϵ -BSVR, DSVR and the proposed robust SVR methods for HBK data set

3.3. Simulation Study

Two simulation studies are conducted to examine the merits of our newly developed robust SVR technique and existing methods (standard SVR and DSVR) in the presence of outliers and high leverage points. The first simulation study deals with the linear case and the second simulation with the nonlinear case. The simulation studies were performed by using R software.

3.3.1. Simulation I: Linear Case

A linear regression model with two independent variables and different sample sizes, that is $n = 50, 100$ and 150 is considered in this simulation study. For each sample $n = 50, 100$ and 150 , the following relationship is used to produce the clean data [24]:

$$y = 1 + 2x_1 + 3x_2 + r_i \quad (24)$$

where x_1 and x_2 are generated from a uniform distribution, $U(-5, 5)$. The errors, r_i are generated from standard normal distribution, $N(0, 1)$. Some clean observations are replaced by contaminated observations in order to create outliers in the y -direction and high leverage points. For each sample, $(\alpha \times n)$ observations for x_1, x_2 and y are replaced by outlying values, where α is the amount of contamination. For high leverage points, the last good observations of x_1 and x_2 are replaced with contaminated observations where $x_1 \sim U(44.5, 50)$ and $x_2 \sim U(50.5, 55)$, holding the corresponding y values the same as the clean data. To create vertical outliers, the third and subsequent good observations are replaced with

$y \sim U(55.5, 60)$, with x_1 and x_2 the same as the clean data. Also, we considered different contamination percentages (10%, 15% and 20%) and different values for the parameters ($\epsilon = 0.1, 0.3$ (as small as possible) and $C = 50, 100$).

For this simulation study, consideration is given ϵ -SVR with a linear kernel. On the basis of (22) and (23) and for each simulation run (1000), MSE and RMSE are obtained on the basis of ϵ -SVR, DSVR and the proposed robust SVR methods. Average MSE and RMSE for 1000 iterations are recorded. This can be used to assess the merits of the methods mentioned above.

The complete results of this simulation study, based on the proposed robust SVR method, DSVR and the ϵ -SVR, are explained graphically in Figures 6, 7 and 8. This shows the estimates of the above techniques for the different sample sizes and levels of contamination. The DSVR successfully detected vertical outliers in this simulation study, based on absolute residuals of SVR. However, it does not detect high leverage points. On the other hand, by using RMD based on MVE, the proposed robust SVR approach detects both the vertical outliers and high leverage points. Additionally, we can clearly see that the proposed robust SVR approach has successfully achieved very low values of MSE and RMSE relative to DSVR and ϵ -SVR for all possible combinations of different sample sizes and different percentages of contamination. This reveals that the proposed robust SVR method is a more efficient technique compared to other approaches, despite the presence of various percentages of contamination points in both x and y directions.

Figure 6. The MSE and RMSE of the ϵ -SVR, the DSVR and the proposed robust SVR methods for different sample sizes and 10% contamination

Figure 7. The MSE and RMSE of the ϵ -SVR, the DSVR and the proposed robust SVR methods for different sample sizes and 15% contamination

Figure 8. The MSE and RMSE of the ϵ -SVR, the DSVR and the proposed robust SVR methods for different sample sizes and 20% contamination

3.3.2. Simulation II: Nonlinear Case

A nonlinear regression model with two independent variables and different sample sizes, which is $n = 50, 100$ and 250 , is considered in this simulation study. For each sample $n = 50, 100$ and 250 , the following relationship is used to generate the clean data [12]:

$$y = \frac{\sin \sqrt{x_1^2 + x_2^2}}{\sqrt{x_1^2 + x_2^2}} + r_i \quad (25)$$

The x_1 and x_2 values are generated with standard uniform distribution, $U(0,1)$. The residuals (additive error) $r_i, i = 1, 2, \dots, n$ are generated on the basis of standard normal distribution, $N(0,1)$. Some good

observations are replaced by contaminated observations in order to generate outliers and high leverage points. Different percentages of contamination, 10%, 15% and 20% are considered. Contamination is achieved by replacing some clean observations of x_1, x_2 and y by contaminated observations where $x_1 \sim U(4.9, 5)$, $x_2 \sim U(29.9, 30)$ and $y \sim U(9.9, 10)$. Also, specific values for the parameters ($\epsilon = 0.1, 0.3$ (as small as possible), $\nu = 0.2, 0.5$ and $\sigma = 1, 5$) are taken into consideration.

In this simulation study, ν -SVR with RBF kernel is considered. The MSE and RMSE based on ν -SVR, DSVR and the proposed robust SVR methods are obtained for each simulation run (1000) on the basis of (22) and (23). The MSE and RMSE averages are recorded for 1000 iterations. This can be used to evaluate the effectiveness

of the techniques suggested.

Figures 9, 10 and 11 present graphically the results of comparison of the proposed robust SVR technique, the DSVR and the ν -SVR of this simulation study. This describes the estimates of the mentioned methods for different sample sizes and levels of contamination. In this simulation study, vertical outliers and high leverages are effectively identified by both DSVR and the proposed robust SVR approach using absolute fitted values and

robust Mahalanobis distance based on MVE respectively. Furthermore, we can obviously see that the proposed robust SVR method has smaller MSE and RMSE than the DSVR and the ν -SVR for all possible combinations of different sample sizes and different percentages of contamination. This demonstrates that the proposed robust SVR approach is the best technique compared to other techniques, in the presence of different levels of contamination.

Figure 9. The MSE and RMSE of the ν -SVR, the DSVR and the proposed robust SVR methods for different sample sizes and 10% contamination

Figure 10. The MSE and RMSE of the ν -SVR, the DSVR and the proposed robust SVR methods for different sample sizes and 15% contamination

Figure 11. The MSE and RMSE of the ν -SVR, the DSVR and the proposed robust SVR methods for different sample sizes and 20% contamination

4. Conclusions

The regular SVR models are good models for estimating the linear and nonlinear relationship of nonparametric regression problems if there are no anomalous data points. However, the approximation of standard SVR models may no longer be efficient due to the possibility of choosing these abnormal observations as support vectors. The DSVR is being proposed as an alternative technique for rectifying the problem. Nonetheless, only ε -SVR with RBF kernel was considered in the DSVR approach. The weights are also based on the sample variance of the fitted values that can still be influenced by outliers. In order to incorporate the DSVR algorithm into all three kinds of SVR models, ε -SVR, ν -SVR and ε -BSVR, with eight different kernel functions and to resolve the limitations of the DSVR procedure, we have proposed a new robust SVR method, based on the MAD and RMD by using MVE approach, to remedy the problems of outliers and high leverage points at the same time with a greater performance by giving a lower weight to them. Two numerical examples and two simulation studies with linear and nonlinear cases based on some of SVR parameters and different sample sizes are employed to investigate the performance of the proposed robust SVR method. MSE and RMSE are used to measure the performances of the proposed robust SVR, DSVR and selected SVR methods. The overall results show that the newly proposed technique performs well as it produces the smallest MSE and RMSE. So, our proposed robust SVR method is a good alternative to deal with contaminated data sets. For future research, we suggest further work on a robust SVR for high dimensional data ($P > n$) and utilizing various loss functions like quadratic, Huber and Tukey loss functions to make a significant contribution to the performance of the estimation process.

Acknowledgements

This research paper is part of MSc final year project under School of Graduate Studies, University Putra Malaysia.

REFERENCES

- [1] C. Cortes, V. Vapnik. Support-vector networks, *Machine learning*, Vol. 20, No. 3, pp. 273-297, 1995.
- [2] H. Yang, K. Huang, L. Chan, I. King, M. R. Lyu. Outliers treatment in support vector regression for financial time series prediction, *International Conference on Neural Information Processing*, Springer, pp. 1260-1265, 2004.
- [3] V. Ceperic, G. Gielen, A. Baric. Sparse ε -tube support vector regression by active learning, *Soft Computing*, Vol. 18, No. 6, pp. 1113 – 1126, 2014.
- [4] A. J. Smola, B. Schölkopf. A tutorial on support vector regression, *Statistics and computing*, Vol. 14, No. 3, pp. 199-222, 2004.
- [5] V. N. Vapnik. The nature of statistical learning theory, *Statistics for engineering and information science*, New York, Vol. 21, pp. 1003-1008, 2000.
- [6] D. Basak, S. Pal, D. C. Patranabis. Support Vector Regression, *Neural Information Processing*, Vol. 11, No. 10, 2007.
- [7] A. Mariette, R. Khanna. Support vector regression, In *Efficient learning machines*, Apress, Berkeley, CA, pp. 67-80, 2015.
- [8] S. Rana, W. Dhhan, H. Midi. Fixed parameters support vector regression for outlier detection, *Economic Computation & Economic Cybernetics Studies & Research*, Vol. 52, No. 2, 2018.
- [9] V. Barnett, T. Lewis. Outliers in statistical data, John Wiley and Sons, New York, 1994.
- [10] P. J. Rousseeuw, B. C. Van Zomeren. Unmasking multivariate outliers and leverage points, *Journal of the American Statistical association*, Vol. 85, No. 411, pp. 633-639, 1990.
- [11] C. C. Chuang, S. F. Su, J. T. Jeng, C. C. Hsiao. Robust support vector regression networks for function approximation with outliers, *IEEE Transactions on Neural Networks*, Vol. 13, No. 6, pp. 1322-1330, 2002.
- [12] W. Dhhan, H. Midi, T. Alameer. Robust Support Vector Regression Model in the Presence of Outliers and Leverage Points, *Modern Applied Science*, Vol. 11, No. 8, 2017.
- [13] E. M. Jordaan, G. F. Smits. Robust outlier detection using SVM regression, Paper presented at the IEEE International Joint Conference on Neural Networks, 2004.
- [14] J. Nishiguchi, C. Kaseda, H. Nakayama, M. Arakawa, Y. Yun. Modified support vector regression in outlier detection. Paper presented at the Neural Networks (IJCNN), The 2010 International Joint Conference on Neural Networks, IEEE, pp. 1-5, 2010.
- [15] W. Dhhan, S. Rana, H. Midi. Non-sparse ε -insensitive support vector regression for outlier detection. *Journal of Applied Statistics*, Vol. 42, No. 8, pp. 1723-1739, 2015.
- [16] B. Schölkopf, A. J. Smola, R. C. Williamson, P. L. Bartlett. New support vector algorithms, *Neural Computation*, Vol. 12, No. 5, pp. 1207–1245, 2000.
- [17] O. L. Mangasarian and D. R. Musicant. Successive overrelaxation for support vector machines, *IEEE Transactions on Neural Networks*, Vol. 10, pp. 1032–1037, 1999.
- [18] C. Leys, C. Ley, O. Klein, P. Bernard, L. Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median, *Journal of Experimental Social Psychology*, Vol. 49, No. 4, pp. 764-766, 2013.
- [19] P. J. Rousseeuw, C. Croux. Alternatives to the median absolute deviation, *Journal of the American Statistical association*, Vol. 88, No. 424, pp. 1273-1283, 1993.

- [20] A. M. Leroy, P. J Rousseeuw. Robust regression and outlier detection, Wiley Series in Probability and Mathematical Statistics, New York, 1987.
- [21] A. Rahmatullah Imon. Identifying multiple influential observations in linear regression, Journal of Applied Statistics, Vol. 32, No. 9, pp. 929-946, 2005.
- [22] T. Hastie, R. Tibshirani, J. Friedman. The elements of statistical learning: prediction, inference and data mining, Springer-Verlag, New York, 2009.
- [23] D. M. Hawkins, D. Bradu, G. V. Kass. Location of several outliers in multiple-regression data using elemental sets, Technometrics, Vol. 26, No. 3, pp. 197-208, 1984.
- [24] H. Midi, L. H. Ann, S. Rana. On the Robust Parameter Estimation for Linear Model with Autocorrelated Errors, Advanced Science Letters, Vol. 19, No. 8, pp. 2494-2496, 2013.