Asian Journal of Chemistry Vol. 22, No. 6 (2010), 4511-4523

Iterative Robust Least Square Support Vector Machine for Spectral Analysis

XIN BAO and LIANKUI DAI*
State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou-310027, Zhejiang, P.R. China
Fax: (86)(571)87952127; Tel: (86)(571)87851894
E-mail: lkdai@iipc.zju.edu.cn

The aim of this study is to develop a novel robust regression algorithm: robust least squares support vector machine (RLS-SVM), to overcome the limitation of the existing support vector machine at high percent of contamination for spectral analysis. In the algorithm, firstly a subset is selected randomly from the original data set to build regression model and the robust estimates of the residuals for the whole set are generated; then the confidence interval of the residuals distribution is applied iteratively to detect outliers. Finally, the LS-SVM estimates are created from the regression model being trained with the selected subset without outliers. The proposed algorithm is applied in the near infrared spectral analysis of gasoline samples in order to predict their octane number with some outliers. Compared with other support vector machine algorithms, the test results show the breakdown point value for the algorithm can be over 45 %. The results also show its priority in predicted precision.

Key Words: Robust regression, Breakdown point, Nonlinear, Least square support vector machine, Spectral analysis.

INTRODUCTION

Spectral analysis combined with chemometrics has proved its efficiency for laboratory and industry applications[1] in providing non-destructive measurement of many chemical properties[2].

Because of nonlinearity between spectra and chemical properties, nonlinear calibration methods perform well in spectral analysis[1]. As one of the nonlinear calibration methods proposed by Vapnik[3], support vector machines (SVM) has become an important novel method in nonlinear calibration due to its remarkable characteristics such as good generalization performance, requiring less training samples, the absence of local minima and the sparse representation of its solutions[4]. In the last decade, the least squares (LS) versions of SVM (LS-SVM) have been investigated[5]. In these LS-SVM formulations one works with equality instead of inequality constraints and a sum of the squared error cost function is used. This reformulation greatly simplifies the problem. SVM and LS-SVM have successfully applied in many areas such as pattern recognition, control systems, signal processing, spectral analysis,

etc.[6-10]. But LS-SVM solutions also have some potential drawbacks such as lost of sparseness. This property might lead to estimates to be less robust.

In the field of spectral analysis, accuracy of the calibration model is most focused on but its robustness is often ignored. However, the main drawback of spectral analysis is linked to its lack of robustness in calibration models when dealing with slight variations in experimental conditions. Samples with extreme characteristics, which called outliers, can also be present in the data, especially in values of chemical properties. Classical calibration methods are strongly influenced by the presence of outliers and the models obtained may not describe the majority of the data well. Thus, robust, not sensitive to the outliers, calibration methods are needed.

Many existed robust calibration methods[11] can only eliminate outliers in spectra. However, they hardly deal with outliers in chemical properties. Some methods are based on linear calibration methods[12-14], so their advantages are not totally display when there is nonlinearity between spectra and properties.

Suykens *et al*.[15] proposed a weighted least squares support vector machine (WLS-SVM) that may provide robust calibration in order to overcome these drawbacks concerning sparseness and robustness in LS-SVM framework. The WLS-SVM works in the case of outliers and tailed non-Gaussian error distributions with good robustness and sparseness. Christmann *et al*.[16] enhanced robustness of support vector machine by revise the influence function, but they do not give the regression results at high per cent of outliers. Chuang *et al*.[17] adopted the concept of traditional robust statistics to fine tune the support vector of support vector machine, simulation results have shown the effectiveness of the approximated function in discriminating against a few outliers. However, it is subjective to determine the proper robust cost function and parameters to iteratively compute and it is needed to specify the percentage of outliers, this is improper for industrial processes. Moreover, its computational cost is huge.

In this paper, we propose a novel robust least squares support vector machine (RLS-SVM) regression algorithm for spectral analysis with good performance at high percent of contamination. In the algorithm, we select the uncontaminated samples as much as possible. In this procedure, a subset is selected randomly from the original data set to estimate the whole set and the robust estimates of the residuals are generated; then the confidence interval of the residuals distribution is applied iteratively to detect outliers, therefore we may eliminate all outliers from the subset. As a result, the LS-SVM estimates are applied to the regression only based on the selected uncontaminated subset.

This paper is organized as follows. First, we give regression model of RLS-SVM and the details in the implementations of the proposed RLS-SVM regression algorithm and give a simulation example. After that, we apply this algorithm in a real world spectral dataset collected from a production scale refinery and present the comparison results with WLS-SVM, SVM and LS-SVM. Finally, the conclusions are addressed.

Theory: LS-SVM can not resist the outliers because of its least squares principle. WLS-SVM is more robust than LS-SVM, however, it can not resist when the per cent of outliers is higher than 25 %. This is because WLS-SVM is also based on least squares. The weighted values are computed based on residuals being computed by LS-SVM. In order to enhance the robustness of excited LS-SVM, we propose a novel robust LS-SVM. We will apply uncontaminated subset for modeling by selecting uncontaminated samples and eliminating outliers through iterative learning. After the uncontaminated subset selected, the LS-SVM regression model based on it is more robust than any other support vector machines.

Let us consider the RLS-SVM regression model. Giving a training dataset with N input/output points $\{x_k, y_k\}^N_{k=1}$ $\{x_k \in R^m$ and $y_k \in R\}$. The new optimization problem can be defined by

$$\min_{w,e,h} J(w,e) = \frac{1}{2} w^T w + \hat{s} \qquad (1)$$
$$\text{s.t. } y_k = w^T j(x_k) + b + e_k, \qquad k = 1, \ldots, N$$

where $\varphi(\cdot)$: $R^m \rightarrow R^{m_h}$ a function which maps the input space R^m into a so-called higher dimensional feature space R^{m_h}, weight vector $w \in R^{m_h}$ is in primal weight space; b is bias term; $\hat{s}$ is the robust estimate of standard deviation[18] for the residuals $\{e_k\}^N_{k=1}$. It can be described as follows:

$$\hat{s} = 1.483 \text{med}\{| e_k - \text{med}\{e_k\}^N_{k=1} |\}^N_{k=1} \qquad (2)$$

In eqn. 2, med means the median of residuals $\{e_k\}^N_{k=1}$. The constant $1.483 = 1/\Phi^{-1}(0.75)$ is an asymptotic correction factor for the case of normal errors[18]. It is applied to adjust $\hat{s}$.

$\hat{s}$ is not disturbed by outliers easily and it can represent most distributions of uncontaminated subset. The eqn. 1 cannot be solved directly. An iterative algorithm will be introduced as follows.

Assume the residuals are normal distribution in most cases without outliers. To the residuals with normal distribution, the confidence interval of residuals[19] can be selected as $[-c \times \hat{s} + \text{med}\{e_k\}^N_{k=1}, c \times \hat{s} + \text{med}\{e_k\}^N_{k=1}]$. The constant c is typically chosen as c = 2.5 for a normal distribution. Because there will be very few residuals larger than $2.5\hat{s}$. So we can choose uncontaminated data belong to the following confidence interval.

$$| (e_k - \text{med}\{e_k\}^N_{k=1})/\hat{s} | \leq c \qquad (3)$$

Based on eqns. 2 and 3, if there are some outliers in e_k, we can find these outliers by getting the robust estimates for residuals and applying confidence interval of residuals distribution. If the per cent of contamination is high, confidence interval will be disturbed by outliers. So we have to repeat this step iteratively. After iteration, we may create the residuals subset without outliers. In the proposed algorithm, we call this step a P-step, where P stands for 'purification' since this step can purify the residuals set by eliminating those outliers in the training dataset. The objective of this step is to select as many uncontaminated samples as possible. The workflow of the P-step can be described as follows.

P-step: Given a residuals set $\{e_k\}^N_{k=1}$ as an input. Let j = 1.

(a) Sort the original dataset $E^j = \{e(k)| k = i_1, \ldots, i_N\}$; (b) Choose the median of E^j and compute $\hat{s}$ the robust estimate of standard deviation for E^j, choose all of the data point $\in [-c \times \hat{s} + \text{med}(e_k), c \times \hat{s} + \text{med}(e_k)]$ and transfer them into $= E^{j+1} = \{e(k)| k = i_1, \ldots, i_{N_1}$, so $E^{j+1} \subseteq E^j$; (c) If $E^{j+1} = E^j$ then break from P-step and get the residuals subset without outliers; (d) Else let $j \leftarrow j + 1$ and go to step (b).

For example, a sorted dataset are giving E^1 = {-15, 1, 2, 3, 4, 5, 6, 15, 20, 25, 30}. At first time, the med (e) is 5, $\hat{s}$ is 5.932 and we choose the data point {1, 2, 3, 4, 5, 6, 15} as E^2; then in the second time, the med (e) is 4, $\hat{s}$ is 2.996 and we choose the data point {1, 2, 3, 4, 5, 6} as E^3; in last time, the med (e) is 3.5, $\hat{s}$ is 2.224 and E^4 = {1, 2, 3, 4, 5, 6} = E^3. So we choose the data subset without outliers.

Our robust algorithm can be divided into three parts: firstly a subset of sample data H with h observations is selected randomly to build a LS-SVM model and estimate all the dataset by this model, so we can get initial residuals subset $\{e_k\}^N_{k=1}$; h means the number of least normal sample[20]. Because the per cent of outliers is not more than 50 %[18], we choose h in the field of $N/2 \leq h \leq N$. Then P-step is used iteratively to select uncontaminated residuals and corresponding samples without outliers. In every iteration, the outliers in subset are discarded and the uncontaminated sample data outside the subset are selected into, so last subset may be contains more than 50 % observations. Finally, we can use the selected training subset to estimate the all set. We can describe our algorithm using a computational work flowchart given in Fig. 1. This procedure is usually repeated several times and we will choose the best answer as the result of RLS-SVM.

This algorithm is inspired by FAST-LTS algorithms[20], but we apply P-step not only to discard outliers but also select the uncontaminated sample data outside of the initial subset and we apply LS-SVM for regression instead of least squares.

EXPERIMENTAL

Data set: Octane number is one of the most important properties of gasoline. It is determined by standard knock intensity in specially designed, ASTM-CFR test engines. The standard measurements are expensive, time-consuming and complicated. From 1989, near infrared (NIR) spectroscopy combined with regression has been extensively used to predict gasoline octane number[21-23].

A group of 250 samples of gasoline was prepared for this experiment, which are scanned by NIR spectra. All the samples were obtained from several refineries in China without additives. The reference value is research octane number (RON), which were measured with ASTM D2699 standard method. The research octane number of these samples distribute from 89-98.

Apparatus and experimental parameters: We measure these spectra by USB-2000 NIR spectrophotometer (Ocean Optics, USA) at 2 nm intervals over a wavelength range of 650-1150 nm. Each sample was scanned 20 times at 50 ms/time and we got the average spectrum. The cell material is quartz and the optical path is 10 cm. The spectra were acquired at room temperature (20-23 °C).

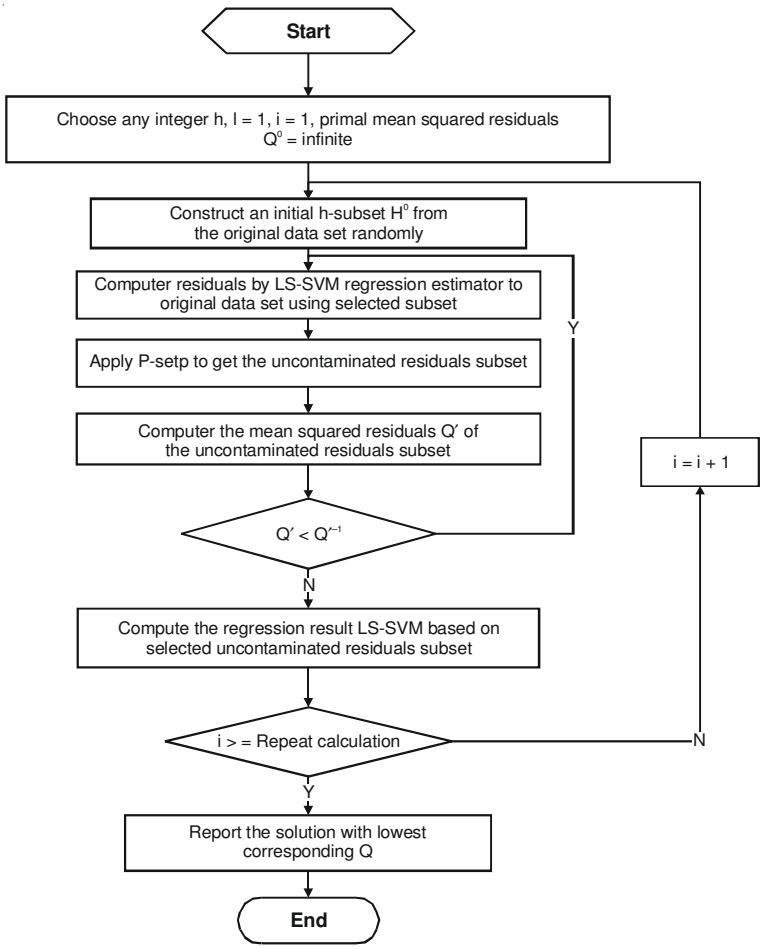

Fig. 1. Computational work flowchart of RLS-SVM

USB-2000 has no thermostatic control. The dark spectrum and the reference spectrum were taken before measuring the sample spectrum. To obtain an analytical signal with satisfactory accuracy, we calculate absorbance spectrum as follows:

$$A = -\lg \frac{I_s - I_d}{I_{ref} - I_d} \tag{4}$$

where I_s means sample spectrum, I_d means dark spectrum and I_{ref} means reference spectrum.

Preprocessing: Standard normal variate (SNV) transformation[24] and Savitzky-Golay smoothing[25] were applied to every absorbance spectrum in order to reduce its fluorescence and improve the signal-noise ratio. The absorbance spectra before and after preprocessing are shown in Fig. 2. In this experiment, we specify the predictor input matrix is the spectral data after preprocessing and the response vector is research octane number.

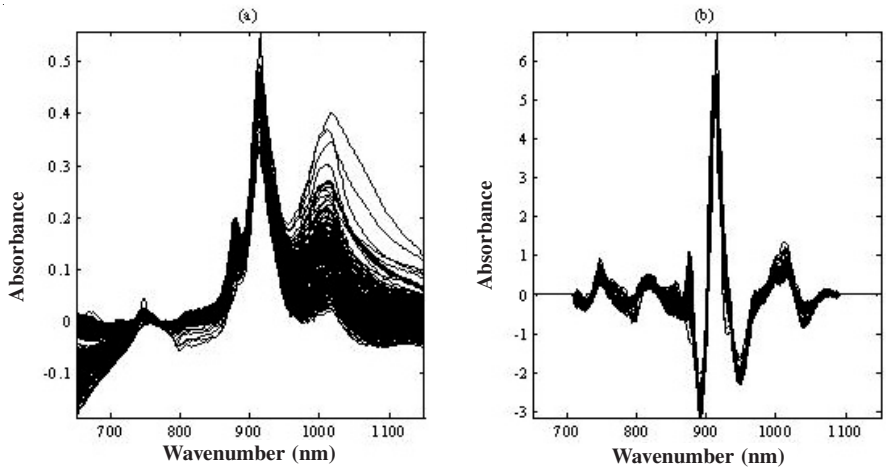

Fig. 2. NIR absorbance spectra of gasoline samples (a) before preprocessing; (b) after preprocessing

Before applying our algorithm, we first apply Mahalanobis distance and Dixon method[26] to detect outliers in spectra, two abnormal spectra can be found out. We finally discard the 2 abnormal samples and select remained 248 samples for modeling.

Modeling and evaluating criterion: We use the RLS-SVM in present experiment compared with WLS-SVM, LS-SVM and standard SVM. All the 248 samples are divided to two parts randomly. To ensure the calibration set and testing set get the same Y-range and data points distribution, we first sort the samples by its research octane number and then select 60 samples as testing set and the remainder 188 samples as calibration set.

Some extra large values are added to research octane number of calibration set randomly in different percentage from 0-50 % to contaminate it. Firstly we testify the robustness of RLS-SVM *via* regression on calibration set, then we use the contaminated calibration set to predict the research octane number in testing set.

To compare the regression performances, an appropriate criterion must be chosen. A robust method should be resistant to outliers in the calibration set. Model robustness can be represented the empirical breakdown point value[27]. Breakdown point value is a percentage; it can be expressed that the estimator will be "breakdown" when the per cent of outliers reaches to some point. Because we apply PCA and mahalanobis distance to detect outliers in spectra above, there are only y-outliers left.

For each percentage of contamination of calibration set, we calculate the contaminated regression value $\hat{y}^c$ and compare it to the original uncontaminated value y^c, let

$$E = \| \hat{y}^c - y^c \| \quad (5)$$

We add different percentage of outliers to the data set from 0-50 % of the data. For each per cent of contamination we calculate the E. For example, if an algorithm

has a breakdown point value of 30 %, E will be low and change little when the per cent of contamination is lower than 30 %. While the per cent of contamination is over 30 %, E would increase dramatically. Therefore, we can use this property to get empirical breakdown point value.

Besides, root mean squared error of prediction (RMSEP) and correlation coefficient (R^2) are considered for predicted precision. Root mean squared error of prediction is defined by eqn. 6 and R^2 is defined by eqn. 7.

$$RMSEP = \sqrt{\frac{1}{N_p}\sum_{i=1}^{N_p}(\hat{y}_i^p - y_i^p)^2} \quad (6)$$

$$R^2 = \sum_{i=1}^{N_p}(\hat{y}_i^p - \overline{y})^2 / \sum_{i=1}^{N_p}(y_i^p - \overline{y})^2 \quad (7)$$

where, N_p is the sample number of the testing set, $\hat{y}^p$ represent the predicted value of RON of testing set, y^p represents the real value of research octane number of testing set and $\overline{y}$ represents the mean of y^p.

Parameters selection: Root mean squared error of prediction is selected as criterion to determine the parameters of the algorithms in calibration estimator. We select the model parameters to minimize the RMSEP. RBF kernel is used in all four algorithms. The optimized parameters are in Table-1.

TABLE-1
OPTIMIZED PARAMETERS

Algorithms	Parameters
RLS-SVM	$\gamma = 20$, $\sigma^2 = 200$, h = n/2, repeated times = 100
WLS-SVM	$\gamma = 20$, $\sigma^2 = 200$, $c_1 = 2.5$, $c_2 = 3$
SVM	$C = 3$, $\varepsilon = 0.001$, $\sigma^2 = 200$
LS-SVM	$\gamma = 20$, $\sigma^2 = 200$

RESULTS AND DISCUSSION

Influence of the constant c: First, we analyzed the influence of the constant c using the following two targets:

$$Diagnosis = \frac{\text{The number of detected outliers}}{\text{The number of all outliers}} \times 100\,\% \quad (8)$$

$$Misdiagnosis = \frac{\text{The number of normal samples detected as outlier}}{\text{The number of all normal samples}} \times 100\,\% \quad (9)$$

We apply RLS-SVM with different value of c on calibration set at different percentage of outliers from 0-50 %. The results can be seen in Fig. 3. When c = 2, the diagnosis keeps 100 % until the per cent of outliers is over 45 %. However, the misdiagnosis is 14 % when there is 5 % of outliers and 9 % when 10 % of outliers. When c = 3, the misdiagnosis keeps 0 % but the diagnosis is near 0 % when the per

cent of outliers just reaches 40 %. Only when c = 2.5, the diagnosis keeps 100 % until the per cent of outliers is over 45 % and the misdiagnosis is no more than 5 %. So we choose c = 2.5 in this paper.

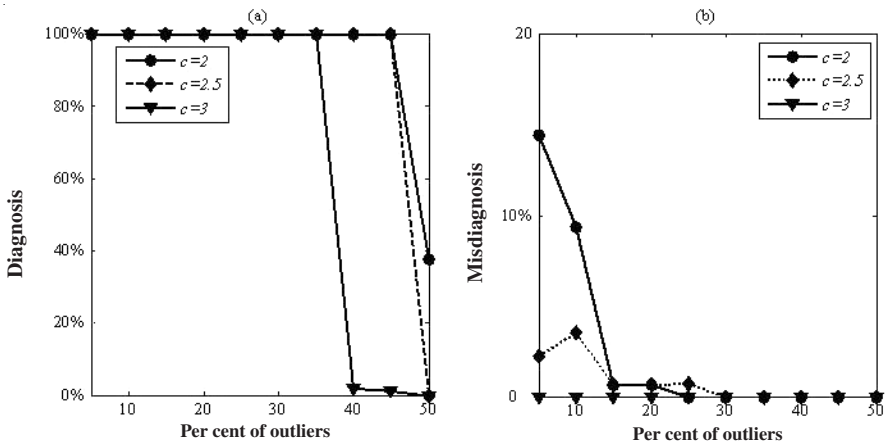

Fig. 3. Diagnosis and misdiagnosis of different values of c (a) diagnosis; (b) misdiagnosis

Robustness: The empirical breakdown point change is shown in Fig. 4. The breakdown point of RLS-SVM in high dimension is about 46 %. The breakdown point of WLS-SVM is still less than 20 %. SVM and LS-SVM can not overcome the influence of large value outliers, so their breakdown point values are very small.

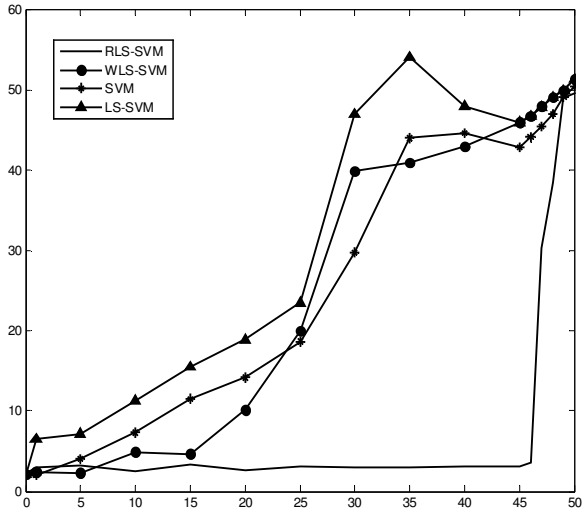

Fig. 4. Empirical breakdown value of four regression algorithms for gasoline data set

We give the value of E at 0, 10 and 20 % of contamination in Table-2. We can see E of RLS-SVM also keeps near 2.5 and change little at these situations.

TABLE-2
E AT 0, 10 AND 20 % OF CONTAMINATION

Algorithms	E		
	0 %	10 %	20 %
RLS-SVM	2.4	2.6	2.7
WLS-SVM	2.4	4.9	10.2
SVM	2.0	7.3	14.2
LS-SVM	2.4	11.3	18.9

The regression estimate scatter diagrams of these algorithms are shown in Fig. 5 when the percentage of contamination is 30 %. RLS-SVM eliminates outliers completely, so its regression results show the distribution of main part of samples and they are close to uncontaminated value. The others are disturbed by outliers. So their regression values are far away from the real values. From Fig. 5, it can be observed that RLS-SVM is much more robust than any other support vector machines.

Predict precision: Now we observe and compare the predicted performances of these four SVM algorithms and a kind of robust PLS algorithm-PLS_OD[28]. Table-3 presents the RMSEP and R^2 of these algorithms at different percentages of contamination.

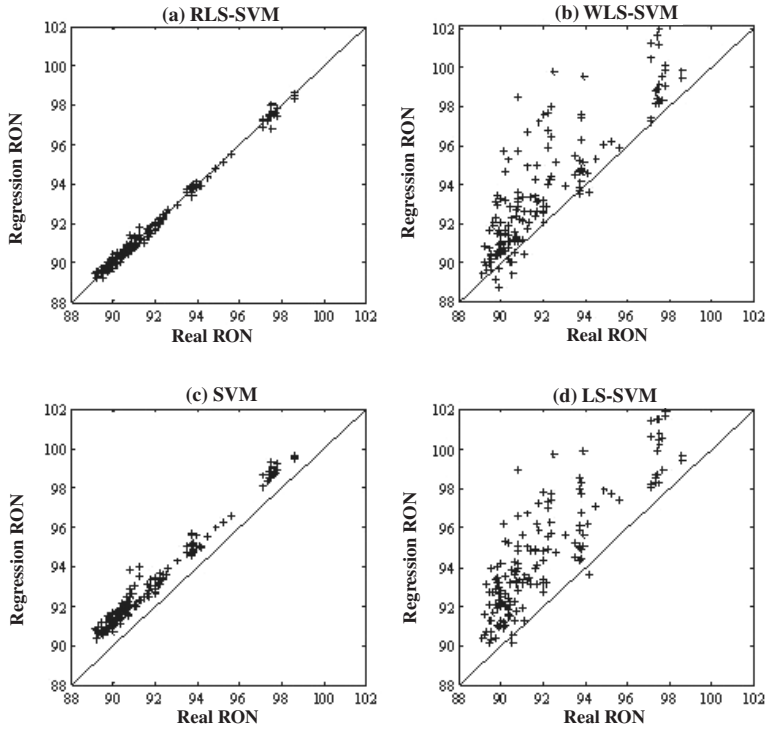

Fig. 5. Scatter diagrams of regression research octane number (RON) *versus* contaminative research octane number at 30 % contamination (a) RLS-SVM; (b) WLS-SVM; (c) SVM: (d) LS-SVM

The results obtained from Table-3 clearly show a fast increase of the RMSEP in LS-SVM and SVM when outliers are present in the training set, whereas RMSEP of RLS-SVM changes little. WLS-SVM do not change obviously at low per cent of outliers, however, when there are 30 % outliers, the RMSEP of them are about 1.93, 4 times as that of RLS-SVM. When there is no outlier, RLS-SVM performs the same as LS-SVM; with the increase of the percentage of contamination, RLS-SVM shows its benefits. PLS_OD has the similar robustness as RLS-SVM, but its predict precision is worse then that of RLS-SVM. For example, RMSEP of PLS_OD with 30 % of contamination is 0.72, almost 1.5 times of RMSEP of RLS-SVM with same contamination. In fact, the RMSEP of RLS-SVM for contaminated data set is still quite comparable to the RMSEP of LS-SVM for the uncontaminated data set.

TABLE-3
PREDICTIVE PERFORMANCES OF FOUR ALGORITHMS
AT DIFFERENT PERCENTAGE OF CONTAMINATION

Per cent of contamination	0 %		10 %		30 %	
Algorithms	RMSEP	R^2	RMSEP	R^2	RMSEP	R^2
RLS-SVM	0.44	0.98	0.47	0.98	0.50	0.97
WLS-SVM	0.43	0.98	0.48	0.98	1.93	0.62
SVM	0.40	0.98	0.65	0.95	1.59	0.74
LS-SVM	0.43	0.98	1.11	0.87	3.06	0.04
PLS_OD	0.69	0.94	0.72	0.93	0.72	0.93

Fig. 6 shows the scatter diagrams of real and predicted research octane number at 30 % of contaminated training set. From Fig. 6, we know only RLS-SVM can predict the testing sample accurately, the others are badly disturbed by the outliers.

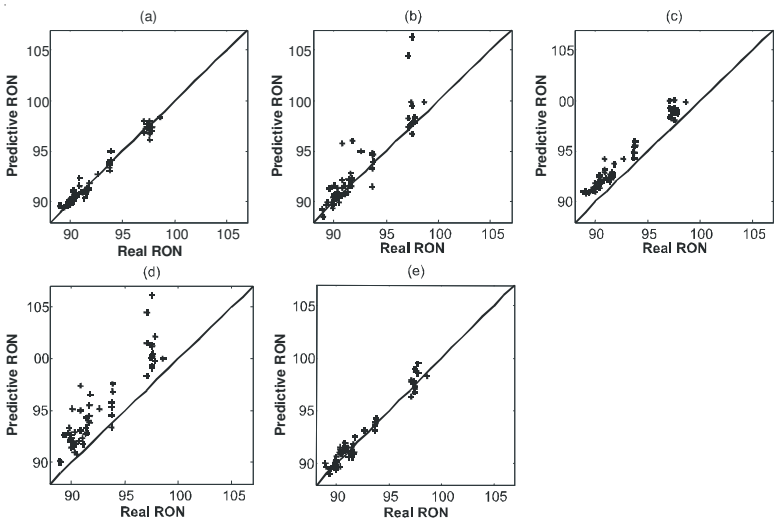

Fig. 6. Scatter diagrams of predicted research octane number (RON) *versus* real research octane number at 30 % contamination (a) RLS-SVM; (b) WLS-SVM; (c) SVM; (d) LS-SVM; (e) PLS_OD

The results obtained by RLS-SVM show a good consistency between predicted and real research octane number of gasoline data sets. In general, it is concluded that the RMSEP of RLS-SVM is much acceptable and the RMSEP for the uncontaminated data set are also comparable.

Influence of different kernel function: Here we compare the influence of different kernel function. Other three kernel functions are used: linear kernel function eqn. 10, polynomial kernel function eqn. 11 and sigmoid kernel function eqn. 12. We still select the parameters to minimize the RMSEP. Therefore the constant q in eqn. 11 is 2 and a and b in eqn. 12 is 1 and -0.75, respectively.

$$K(x_i, x_j) = x_i \cdot x_j \quad (10)$$

$$K(x_i, x_j) = [(x_i \cdot x_j) + 1]^q \quad (11)$$

$$K(x_i, x_j) = \tan h\,[a(x_i, x_j) + b] \quad (12)$$

We first observe the breakdown point of these RLS-SVMs. It can be seen in Fig. 7. Linear kernel function and polynomial kernel function are unsuitable for spectral analysis and breakdown points of RLS-SVM with these two functions are very low, not more than 15 %. Sigmoid kernel function is nonlinear function; it is suitable for this experiment, so the breakdown point of RLS-SVM with sigmoid kernel function is similar with that of RBF kernel function. But its regression residuals are larger than that of RBF kernel function.

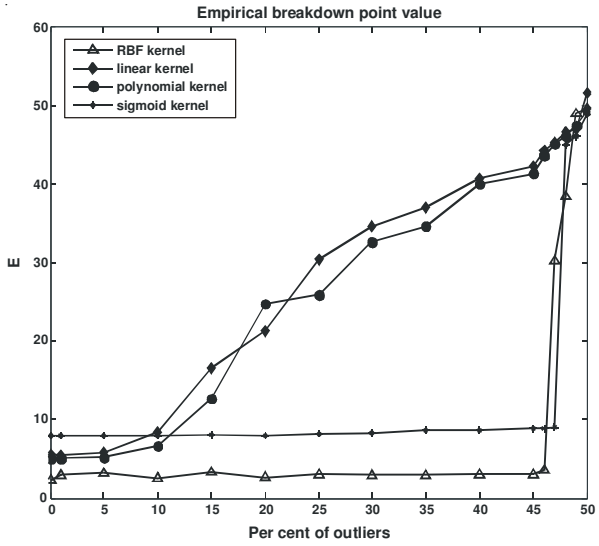

Fig. 7. Empirical breakdown value of RLS-SVM with four kernel functions for gasoline data set

Second, we observe the predicted performances of these RLS-SVMs. The training set with 10 % outliers is still used to predict the test set. The results are shown in Table-4. The predicted result of RLS-SVM with RBF kernel function is the best. So RBF kernel function is a reasonable choice.

TABLE-4
PREDICTIVE PERFORMANCES OF RLS-SVM WITH DIFFERENT KERNEL FUNCTION AT 10 % OF OUTLIERS

Kernel function	RMSEP	R^2
RBF	0.50	0.97
Linear	0.99	0.90
Polynomial	0.97	0.90
Sigmoid	0.68	0.95

Conclusion

While SVM and LS-SVM have been widely applied in spectral analysis, however, they are not robust when the y-outliers exist in the training dataset. WLS-SVM can overcome the influence of outliers only when the percentage of contamination is less then 20 % in our experiments. Nevertheless, the robustness can be further enhanced by our RLS-SVM. The experimental results have shown that it is robust towards contamination, whereas its performance is also good for uncontaminated data sets. The breakdown point of RLS-SVM exceeds 45 %. RLS-SVM algorithms proved their superiority over other LS-SVM algorithms both in robustness and predicted accuracy.

ACKNOWLEDGEMENT

This work was supported by the National High Technology Research and Development Program ("863"Program) of China (Grant No. 2006AA040309).

REFERENCES

1. F. Chauchard, R. Cogdill, S. Roussel, J.M. Roger and V. Bellon-Maurel, *Chemom. Intell. Lab. Syst*., **71**, 141 (2004).
2. B. Osborne, T. Fearn and P.H. Hindle, Near Infrared Spectroscopy in Food Analysis, John Wiley and Sons, New York (1986).
3. V. Vapnik, Statistical Learning Theory, John Wiley & Sons, New York (1986).
4. J.A.K. Suykens and J. Vandewalle, Nonlinear Modeling: Advanced Black-box Techniques, Kluwer Academic Publishers, Boston (1998).
5. J.A.K. Suykens and J. Vandewalle. *Neural Process. Lett*., **9**, 293 (1999).
6. Y. Zhang, Q. Xiong, G. Yang, M.L. Li and J. Zhang, *Anal. Sci*., **23**, 911 (2007).
7. A. Niazi, J. Zolgharnein and S. Afiuni-Zadeh, *Anal. Sci*., **23**, 1311 (2007).
8. C. Tan, M.L. Li and X. Qin, *Anal. Sci*., **24**, 647 (2008).
9. A. Borin, M.F. Ferrao, C. Mello, D.A. Maretto and R.J. Poppi, *Anal. Chim. Acta*, **579**, 25 (2006).
10. F. Chauchard, J. Svensson, J. Axelsson, S. Andersson-Engels and S. Roussel, *Chemom. Intell. Lab. Syst*., **91**, 34 (2008).
11. W.J. Egan and S.L. Morgan, *Anal. Chem*., **79**, 2372 (1998).
12. B. Walczak and D.L. Massart, *Chemom. Intell. Lab. Sys*., **27**, 41 (1995).
13. K.V. Branden and M. Hubert, *Anal. Chim. Acta*, **515**, 229 (2004).
14. S. Serneels, C. Croux, R. Filzmoser and P.J. Van Espen, *Chemom. Intell. Lab. Sys*., **79**, 55 (2005).
15. J.A.K. Suykens, J. de Brabanter, L. Lukas and J. Vandewalle, *Neurocomputing*, **48**, 85 (2002).
16. A. Christmann and I. Steinwart, *J. Mach. Learn. Res*., **5**, 1007 (2004).
17. C. Chuang, S. Su, J. Jeng and C. Hsiao, *IEEE. T. Neural. Networ*., **13**, 1322 (2002).

18. P.J. Rousseeuw and A.M. Leroy, Robust Regression and Outlier Detection, John Wiley & Sons, New York (2003).
19. H.A. David, *Stat. Sci*., **13**, 368 (1998).
20. P.J. Rousseeuw and V. Driessen, *Data. Min. Knowl. Disc*, **12**, 29 (2006).
21. J.J. Kelly, C.H. Barlow, T.M. Jinguji and J.B. Callis, *Anal. Chem*., **61**, 313(1989).
22. R.M. Balabin, R.Z. Safieva and E.I. Lomakina, *Chemom. Intell. Lab. Sys*., **88**, 183(2004).
23. K. Brudzewski, A. Kesik, K. Kolodziejczyk, U. Zborowska and J. Ulaczyk, *Fuel*, **85**, 553 (2006).
24. R.J. Barnes, M.S. Dhanoa and S.J. Lister, *Appl. Spectrosc*., **43**, 772 (1989).
25. A. Savitzky and M.J.E. Golay, *Anal. Chem*., **36**, 1627 (1964).
26. H.Y. Yu, Y.B. Ying, X.P. Fu, H.S. Lu, *J. Near Infrared. Spec*., **14**, 37 (2006).
27. K.V. Branden and M. Hubert, *Anal. Chim. Acta*, **515**, 229 (2004).
28. X. Bao and L. Dai, *Fuel*, **88**, 1216 (2009).

(*Received*: 4 September 2009; *Accepted*: 15 February 2010) AJC-8438