2012 Fifth International Joint Conference on Computational Sciences and Optimization

Parameters selection and noise estimation of SVM regression

Jifu Nong
College of Science
Guangxi University for Nationalities
Nanning, China
njf93471@163.com

***Abstract*—We investigate practical selection of hyper-parameters for support vector machines (SVM) regression. The proposed methodology advocates analytic parameter selection directly from the training data, rather than re-sampling approaches commonly used in SVM applications. In particular, we describe a new analytical prescription for setting the value of insensitive zone, as a function of training sample size. Good generalization performance of the proposed parameter selection is demonstrated empirically using several low-dimensional and high-dimensional regression problems. Further, we point out the importance of Vapnik insensitive loss for regression problems with finite samples. To this end, we compare generalization performance of SVM regression with regression using least-modulus loss and standard squared loss. These comparisons indicate superior generalization performance of SVM regression under sparse sample settings, for various types of additive noise.**

***Keywords*-loss function; parameter selection; prediction accuracy; support vector machine regression**

I. Introduction

This study is motivated by a growing popularity of support vector machines (SVM) for regression problems. Their practical success can be attributed to solid theoretical foundations based on VC-theory, since SVM generalization performance does not depend on the dimensionality of the input space. However, many SVM regression application studies are performed by expert users. Since the quality of SVM models depends on a proper setting of SVM hyper-parameters, the main issue for practitioners trying to apply SVM regression is how to set these parameter values for a given data set. Whereas existing sources on SVM regression give some recommendations on appropriate setting of SVM parameters, there is no general consensus and many contradictory opinions. Hence, re-sampling remains the method of choice for many applications. Unfortunately, using re-sampling for tuning several SVM regression parameters is very expensive in terms of computational costs and data requirements.

This paper describes simple yet practical analytical approach to SVM regression parameter setting directly from the training data. Proposed approach is based on well-known theoretical understanding of SVM regression that provides the basic analytical form of proposed prescriptions for parameter selection. Further, we perform empirical tuning of these analytical dependencies using synthetic data sets. Practical validity of the proposed approach is demonstrated using several low and high-dimensional regression problems.

II. Support vector regression and SVM parameter selection

We consider standard regression formulation under general setting for predictive learning. The goal is to estimate unknown real-valued function in the relationship:

$$y = r(\mathrm{x}) + \delta \tag{1}$$

where δ is independent and identically distributed zero mean random error (noise), x is a multivariate input and y is a scalar output. The estimation is made based on a finite number of samples: $(\mathrm{x}_i, y_i), i = 1, 2, \cdots, n$ The training data are i.i.d. samples generated according to some joint probability density function

$$p(\mathrm{x}, y) = p(\mathrm{x})p(y|\mathrm{x}) \tag{2}$$

The unknown function in Eq. (1) is the mean of the output conditional probability

$$r(\mathrm{x}) = \int yp(y|\mathrm{x})dy \tag{3}$$

A learning method (or estimation procedure) selects the best model $f(\mathrm{x}, \omega_0)$ from a set of approximating functions $f(\mathrm{x}, \omega)$ parameterized by a set of parameters $\omega \in \Omega$. The quality of an approximation is measured by the loss or discrepancy measure $L(y, f(\mathrm{x}, \omega))$, and the goal of learning is to select the best model minimizing (unknown) prediction risk:

$$R(\omega) = \int L(y, f(\mathrm{x}, \omega))p(\mathrm{x}, y)d\mathrm{x}dy \tag{4}$$

It is known that the regression function (3) is the one minimizing prediction risk (4) with the squared loss function loss:

$$L(y, f(\mathrm{x}, \omega)) = (y - f(\mathrm{x}, \omega))^2 \tag{5}$$

Note that the set of functions $f(\mathrm{x}, \omega), \omega \in \Omega$ supported by a learning method may or may not contain the regression function (3). Thus, the problem of regression estimation is

978-0-7695-4690-2/12 $26.00 © 2012 IEEE
DOI 10.1109/CSO.2012.33

the problem of finding the function $f(\mathrm{x}, \omega_0)$ (regressor) that minimizes the prediction risk functional

$$R(\omega) = \int (y - f(\mathrm{x}, \omega))^2 p(\mathrm{x}, y) d\mathrm{x} dy \tag{6}$$

using only the training data. This risk functional measures the accuracy of the learning methods predictions of unknown target function $r(\mathrm{x})$.

In SVM regression, the input x is first mapped onto an m-dimensional feature space using some fixed (non-linear) mapping, and then a linear model is constructed in this feature space . Using mathematical notation, the linear model (in the feature space) $f(\mathrm{x}, \omega_0)$ is given by

$$f(\mathrm{x}, \omega_0) = \sum_{j=1}^{m} \omega_j g_j(\mathrm{x}) + b \tag{7}$$

where $g_j(\mathrm{x}), j = 1, 2, \cdots, m$ denotes a set of non-linear transformations, and b is the bias term.

Regression estimates can be obtained by minimization of the empirical risk on the training data. Typical loss functions used for minimization of empirical risk include squared error and absolute value error. SVM regression uses a new type of loss function called 1-insensitive loss proposed by Vapnik (1998, 1999):

$$L_\varepsilon(y, f(\mathrm{x}, \omega)) = \begin{cases} 0, \text{ if } |y - f(\mathrm{x}, \omega)| \le \varepsilon \\ |y - f(\mathrm{x}, \omega)| - \varepsilon, \text{ otherwise} \end{cases} \tag{8}$$

The empirical risk is:

$$R_{\text{emp}}(\omega) = \frac{1}{n} \sum_{i=1}^{n} L_\varepsilon(y_i, f(\mathrm{x}_i, \omega)) \tag{9}$$

SVM regression performs linear regression in the high dimensional feature space using ε-insensitive loss and, at the same time, tries to reduce model complexity by minimizing $||\omega||^2$. This can be described by introducing (non-negative) slack variables $\xi, \xi_i^*, i = 1, 2, \cdots, n$ to measure the deviation of training samples outside ε-insensitive zone. Thus, SVM regression is formulated as minimization of the following functional:

$$\text{minimize} \quad \frac{1}{2}||\omega||^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)$$

$$\text{subject to} \quad \begin{cases} y_i - f(\mathrm{x}_i, \omega) - b \le \varepsilon + \xi_i^* \\ f(\mathrm{x}_i, \omega) + b - y_i \le \varepsilon + \xi_i \\ \xi_i \ge 0, \xi_i^* \ge 0 \end{cases} \tag{10}$$

where C is a positive constant (regularization parameter). This optimization formulation can be transformed into the dual problem (Vapnik, 1998, 1999), and its solution is given by

$$f(\mathrm{x}) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(\mathrm{x}_i, \mathrm{x}) + b \tag{11}$$

where the dual variables are subject to constraints $0 \le \alpha_i, \alpha_i^* \le C$ and the kernel function $K(\mathrm{x}, \mathrm{x}')$ is a symmetric function satisfying Mercer's conditions (Vapnik, 1998, 1999). The sample points that appear with non-zero coefficients in Eq.(11) are called support vectors (SVs).

III. Proposed approach for parameter selection

Selection of parameter C. Following Mattera and Haykin (1999), consider standard parameterization of SVM solution given by Eq. (11), assuming that the ε-insensitive zone parameter has been chosen. Also suppose, without loss of generality, that the SVM kernel function is bounded in the input domain. For example, RBF kernels satisfy this assumption:

$$K(\mathrm{x}_i, \mathrm{x}) = \exp\left(-\frac{||\mathrm{x} - \mathrm{x}_i||^2}{2p^2}\right) \tag{12}$$

where p is the width parameter.

Under these assumptions, one can relate the value of C to the range on response values of the training data. Specifically, referring to Eq. (11), note that the regularization parameter C defines the range of values $0 \le \alpha_i, \alpha_i^* \le C$ assumed by dual variables used as linear coefficients in SVM solution (11). Hence, a good value for C can be chosen equal to the range of output (response) values of training data. However, such a selection of C is quite sensitive to possible outliers (in the training data), so we propose instead the following prescription for regularization parameter:

$$C = \max\{|\bar{y} + 3\sigma_y|, |\bar{y} - 3\sigma_y|\} \tag{13}$$

where $\bar{y}$ and σ_y are the mean and the standard deviation of the y values of training data.

Selection of ε. First, let us try to relate the value of 1 to an empirical distribution of errors $\delta_i = \hat{y}_i - y_i, i = 1, 2, \cdots, n$ observed for a given training data set of size n: Consider the sample mean of these errors:

$$\hat{\delta} = \frac{1}{n} \sum_{i=1}^{n} \delta_i \tag{14}$$

Random variable $\hat{\delta}$ can be interpreted as empirical estimate of noise observed (or derived) from available training data set of size n. Hence, the choice of ε should depend on the variance of $\hat{\delta}$. In order to estimate the variance of $\hat{\delta}$, recall that component errors δ_i in expression (14) all have zero mean and variance σ^2 (where σ^2 is the variance of additive noise in regression formulation (1)). According to the Central Limit Theorem, the sample mean (14) is (approximately) Gaussian with zero mean and variance σ^2/n. Hence, it seems reasonable to set the value of ε proportional to the width of the distribution of $\hat{\delta}$.

$$\varepsilon \sim \frac{\sigma}{\sqrt{n}} \tag{15}$$

Based on a number of empirical comparisons, we found that Eq. (15) works well when the number of samples is small, however, for large values of n prescription (15) yields ε-values that are too small. Hence, we propose the following (empirical) dependency:

$$\varepsilon \sim \sigma\sqrt{\frac{\ln n}{n}} \tag{16}$$

We do not have specific theoretical justification for factor ln n in the above expression, other that this factor typically appears in analytical bounds used in VC theory (Vapnik, 2001). Based on the empirical tuning, we found the following practical prescription for ε:

$$\varepsilon = 3\sigma\sqrt{\frac{\ln n}{n}} \tag{17}$$

This expression provides good performance for various data set sizes, noise levels and target functions for SVM regression. Expression (17) will be used in all empirical comparisons presented in Sections 4 and 5.

IV. Experimental results for non-linear target functions

This section presents empirical comparisons for nonlinear regression, first with Gaussian noise, and then with non-Gaussian noise.

A. Results for Gaussian noise

First, we describe the experimental procedure used for comparisons, and then present the empirical results.

Training data. Simulated training data $(\mathrm{x}_i, y_i), i = 1, \cdots, n$, where x-values are sampled on uniformly spaced grid in the input space, and y-values are generated according to statistical model (1), i.e. $y = r(\mathrm{x} + \delta)$. Different types of the target functions $r(\mathrm{x})$ are used. The y-values of training data are corrupted by additive noise d with zero mean and standard deviation σ.

Test data. The test inputs are sampled randomly according to uniform distribution in x-space.

Performance metric. Prediction risk is defined as the mean squared error (MSE) between SVM estimates and the true values of the target function for test inputs.

The first set of results show how SVM generalization performance depends on the proper choice of SVM parameters for univariate $sinc$ target function:

$$r(x) = a\frac{\sin x}{x}, x \in [-10, 10] \tag{18}$$

The following values of $a = 1, 10, 0.1, -10, -0.1$, were used to generate five data sets using small sample size ($n = 30$) with additive Gaussian noise (with different noise levels σ shown in Table 1). For these data sets, we used RBF kernels with width parameter $p = 3$. Table 1 shows:

(a) Parameter values C and ε (using expressions proposed in Section 3) for different training sets.

Table I
Results for univariate sinc function

Data Set	a	σ	C	ε	Prediction risk	% SV
1	1	0.2	1.58	0	0.0129	100
				0.2	0.0065	43.3
2	10	2	15	0	1.3043	100
				2.0	0.7053	36.7
3	0.1	0.02	0.16	0	0.000103	100
				0.02	0.00000805	40
4	-10	0.2	14.9	0	0.0317	100
				0.2	0.0265	50
5	-0.1	0.02	0.17	0	0.00014	100
				0.02	0.000101	46.7

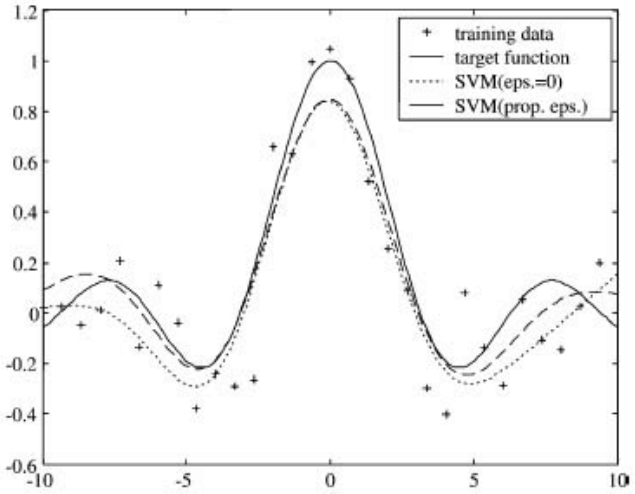

Figure 1. Comparison of SVM estimate using proposed parameter selection versus using least-modulus loss, for Data Set 1 (sinc target function, 30 samples).

(b) Prediction risk and percentage of support vectors (% SV) obtained by SVM regression with proposed parameter values.

(c) Prediction risk and % SV obtained using LM loss function ($\varepsilon = 0$).

We can see that the proposed method for choosing ε is better than LM loss function, as it yields lower prediction risk and better (more sparse) representation.

Visual comparisons (for univariate $sinc$ function, Data Set 1) between SVM estimates using proposed parameter selection and using LM loss are shown in Fig.1, where the solid line is the target function, the '+' denotes training data, the dotted line is an estimate using LM loss and the dashed line is the SVM estimate using proposed parameter settings.

Next we show results of SVM parameter selection for multivariate regression problems. The first data set is generated using two-dimensional $sinc$ target function.

$$r(\mathrm{x}) = \frac{\sin\sqrt{x_1^2 + x_2^2}}{\sqrt{x_1^2 + x_2^2}} \tag{19}$$

Table II
COMPARISON OF THE PROPOSED METHOD FOR ε-SELECTION WITH LEAST-MODULUS LOSS $\varepsilon = 0$ FOR TWO-DIMENSIONAL $sinc$ TARGET FUNCTION DATA SETS

Noise level	ε-selection	Prediction risk	%SV
$\sigma = 0.1$	0	0.008	100
	Proposed	0.002	62.7
$\sigma = 0.4$	0	0.0369	100
	Proposed	0.0229	60.9

Table III
COMPARISON OF THE PROPOSED METHOD FOR ε-SELECTION WITH LEAST-MODULUS LOSS $\varepsilon = 0$ FOR HIGH-DIMENSIONAL ADDITIVE TARGET FUNCTION

Noise level	ε-selection	Prediction risk	%SV
$\sigma = 0.1$	0	0.0443	100
	Proposed	0.0387	86.7
$\sigma = 0.2$	0	0.1071	100
	Proposed	0.0918	90.5

defined on a uniform square lattice $[25, 5] \times [25, 5]$, with response values corrupted with Gaussian noise ($\sigma = 0.1$ and $\sigma = 0.4$, respectively). The number of training samples is 169, and the number of test samples is 676. The RBF kernel width parameter $p = 2$ is used. The proposed approach selects the following values $C = 1.16$ and $\varepsilon = 0.05$ (for $\sigma = 0.1$) and $\varepsilon = 0.21$ (for $\sigma = 0.4$). Table 2 compares SVM estimates (with proposed parameter selection) and estimates obtained using LM loss, in terms of prediction risk and the percentage of SV chosen by each method.

Finally, consider higher dimensional additive target function

$$r(\mathrm{x}) = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 \quad (20)$$

where x-values are distributed in hypercube $[0, 1]^5$. Output (response) values of training samples are corrupted by additive Gaussian noise (with $\sigma = 0.1$ and $\sigma = 0.2$). Training data size is $n = 243$ samples (i.e. 3 points per each input dimension). The test size is 1024. The RBF kernel width parameter $p = 0.8$ is used for this data set. The proposed method yields the value of $C = 34$ and the value of $\varepsilon = 0.045$ for $\sigma = 0.1$ and $\varepsilon = 0.09$ for $\sigma = 0.2$. Comparison results between the proposed methods for parameter selection with the method using LM loss function are shown in Table 3. Clearly, the proposed approach gives better performance in terms of prediction risk and robustness.

B. Results for non-Gaussian noise

Next we present empirical results for regression problems with non-Gaussian additive symmetric noise in the statistical model (1). The main motivation is to demonstrate practical advantages of Vapnik' ε-insensitive loss versus other (robust) loss functions. Specifically, we perform empirical comparisons between SVM regression (with proposed parameter selection) versus SVM regression using LM loss $\varepsilon = 0$ for several finite sample regression problems.

Table IV
COMPARISON RESULTS OF SVM WITH PROPOSED ε- SELECTION VERSUS LEAST-MODULUS LOSS $\varepsilon = 0$ FOR t-DISTRIBUTION OF NOISE

Noise level	ε-selection	Prediction risk
$\sigma = 0.1$	0	0.03
	Proposed	0.003
$\sigma = 0.2$	0	0.015
	Proposed	0.014
$\sigma = 0.3$	0	0.031
	Proposed	0.029

Table V
COMPARISON RESULTS OF SVM WITH PROPOSED ε- SELECTION VERSUS LEAST-MODULUS LOSS $\varepsilon = 0$ FOR UNIFORM NOISE

Noise level	ε-selection	Prediction risk
$\sigma = 0.1$	0	0.005
	Proposed	0.004
$\sigma = 0.2$	0	0.020
	Proposed	0.013
$\sigma = 0.3$	0	0.042
	Proposed	0.022

We consider three types of non-Gaussian noise

- Student's t-distribution noise
- Uniform distributed noise
- Laplacian noise.

Univariate $sinc$ target function is used for comparisons:

$$r(x) = \frac{\sin x}{x}, x \in [-10, 10]$$

Training sample size $n = 30$. The x values are sampled on a uniformly spaced grid in the input space. RBF kernels with width parameter $p = 3$ are used for this data set. According to proposed expressions (13) and (17), $C = 1.6$, $\varepsilon = 0.1$ (for $\sigma = 0.1$), $\varepsilon = 0.2$ (for $\sigma = 0.2$), $\varepsilon = 0.3$ (for $\sigma = 0.3$). The comparison results show prediction risk obtained using SVM regression and using LM loss, on the same data sets. In order to perform more meaningful comparisons, all comparison results are averaged using 100 random realizations of the training data.

First, consider Student's t-distribution for noise. Several experiments have been performed using various degrees of freedom (DOF) (40, 50, 100) for generating t-distribution. Empirical results indicate superior performance of the proposed method for SVM parameter selection, in comparison with LM loss regression. Table 4 shows comparisons with regression estimates obtained using LM loss for Student's noise (with 100 DOF) for different noise levels σ.

Second, consider uniform distribution for the additive noise. Table 5 shows comparison results for different noise levels σ. These results indicate superior performance of SVM method with proposed selection of ε.

Finally, we show comparison results for Laplacian noise density. Smola et al. (1998) suggest that for this noise density model, the LM loss should be used. We compare the

Table VI
COMPARISON RESULTS OF SVM WITH PROPOSED ε- SELECTION VERSUS LEAST-MODULUS LOSS $\varepsilon = 0$ FOR LAPLACIAN NOISE

Noise level	ε-selection	Prediction risk
$\sigma = 0.1$	0	0.003
	Proposed	0.004
$\sigma = 0.2$	0	0.010
	Proposed	0.015
$\sigma = 0.3$	0	0.019
	Proposed	0.030

proposed approach for choosing ε with the LM loss method. Empirical results in Table 6 indicate that for this data set, the LM loss $\varepsilon = 0$ yields better prediction accuracy than SVM loss with proposed parameter selection, in agreement with Smola et al. (1998).

V. NOISE VARIANCE ESTIMATION

The proposed method for selecting ε relies on the knowledge of the standard deviation of noise σ. The problem, of course, is that the noise variance is not known a priori, and it needs to be estimated from training data $(\mathrm{x}_i, y_i), i = 1, 2, \cdots, n$.

In practice, the noise variance can be readily estimated from the squared sum of residuals of the training data. Namely, the well-known approach of estimating noise variance is by fitting the data using low bias model and applying the following formula to estimate noise

$$\hat{\sigma}^2 = \frac{n}{n-d} \cdot \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{21}$$

where d is the 'degrees of freedom' (DOF) of the high complexity estimator and n is the number of training samples.

We used expression (21) for estimating noise variance using higher-order algebraic polynomials and k-nearest-neighbors regression. Both approaches yield very accurate estimates of the noise variance; however, we only show the results of noise estimation using k-nearest-neighbors regression. In k-nearest-neighbors method, the function is estimated by taking a local average of the training data. Locality is defined in terms of the k data points nearest the estimation point. Accurate estimates of the model complexity (DOF) for k-nearest neighbors are not known, even though an estimate $d = n/k$ is commonly used. Cherkassky and Ma (2003) recently introduced new estimate of model complexity:

$$d = n/(n^{1/5}k) \tag{22}$$

This estimate of DOF for k-nearest-neighbors regression provides rather accurate noise estimates when used in conjunction with Eq.(22). Combining expressions (22) and (23), we obtain the following prescription for noise variance estimation via k-nearest-neighbor's method.

$$\hat{\sigma}^2 = \frac{n^{1/5}k}{n^{1/5}k - 1} \cdot \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{23}$$

VI. CONCLUSION

This paper describes practical recommendations for setting meta-parameters for SVM regression. Namely the values of ε and C parameters are obtained directly from the training data and (estimated) noise level. Extensive empirical comparisons suggest that the proposed parameter selection yields good generalization performance of SVM estimates under different noise levels, types of noise, target functions and sample sizes. Hence, the proposed approach for SVM parameter selection can be immediately used by practitioners interested in applying SVM to various application domains.

Our empirical results suggest that with the proposed choice of ε, the value of regularization parameter C has only negligible effect on the generalization performance (as long as C is larger than a certain threshold determined analytically from the training data). The proposed value of C-parameter is derived for RBF kernels; however, the same approach can be applied to other kernels bounded in the input domain. For example, we successfully applied proposed parameter selection for SVM regression with polynomial kernel defined in [0,1] (or [21,1]) input domain.

REFERENCES

[1] Chapelle, O., Vapnik, V, *Model selection for support vector machines*. Advances in neural information processing systems 12, 311C321, 1999.

[2] Cherkassky, V., Ma, Y, *Comparison of model selection for regression*. Neural Computation, 15(7), 1691C1714, 2003.

[3] Cherkassky, V., Shao, X., Mulier, F., Vapnik, V, *Model complexity control for regression using VC generalization bounds*. IEEE Transaction on Neural Networks, 10(5), 1075C1089. 1999.

[4] Hastie, T., Tibshirani, R., Friedman, J, *The elements of statistical learning: Data mining, inference and prediction*. Berlin: Springer. 2001.

[5] Mattera, D., Haykin, S, *Support vector machines for dynamic reconstruction of a chaotic system*. Cambridge, MA: MIT Press, 1999.

[6] Muller, K., Smola, A., Ratsch, G., *Using support vector machines for time series prediction*. Cambridge, MA: MIT Press, 1999.

[7] Scholkopf, B., Burges, J., Smola, A, *Advances in kernel methods: Support vector machine*. Cambridge, MA: MIT Press, 1999.

[8] Smola, A., Murata, N, *Asymptotically optimal choice of ε-loss for support vector machines*. Proceedings of ICANN, 1998.

[9] Smola, A., Scholkopf, *A tutorial on support vector regression*. Royal Holloway College, University of London, UK, 1998.

[10] Vapnik, V, *The nature of statistical learning theory (2nd ed)*. Berlin: Springer, 1999.