Selection of Support Vector Machines Parameters for Regression Using Nested Grids

Alexander Popov*, Alexander Sautin*
* NSTU/Department of Software and Database Engineering, Novosibirsk, Russia

Abstract— **The paper examines support vector machines for regression problem. Analysis of different grid types for selection of SVM parameters is conducted. Experimental results obtained on the basis of applying nested grids are presented, and efficiency of regression problem solution with various support vector machine parameters values is investigated.**

I. Introduction

Support vector machines (SVM) [1] are one of the most promising approaches based on the learning for regression. In the formulation of SVM the model includes some hyper-parameters such as the kernel parameter and the regularization parameter that control the generalization performance of SVM. If one is using arbitrary SVM parameters the performance of SVM could be vary in a wide range. Finding the hyper-parameters with a good generalization performance is crucial for the successful application of SVM.

There are different methods of SVM parameters selection but none of them is universal. In each case one has to decide which method of parameters selection he wants to use. In this article we will compare methods of parameters selection which are based on a grid search method and heuristic approach.

II. Support Vector Machines

The aim of regression problem is to construct unknown function in equation

$$y_i = r(x_i) + e_i ,$$

where $x_i \in X \subseteq R^d$ are sample points, $y_i \in Y \subseteq R$ is a response observed at x_i, e_i is error. Function estimation is based on the training data set $(x_1, y_1), \ldots, (x_n, y_n) \in X \times Y$.

In support vector machines, the nonlinear regression problem in the input space X is considered as a linear one $f(x) = w^T \phi(x) + b$ in feature space F. This feature space F is induced by the nonlinear mapping $\phi : x \to \phi(x)$. SVM regression is formulated as minimization of the following functional

$$\min_{w,b,\xi,\xi^*} \left[\frac{1}{2} w^T w + C \sum_{k=1}^{n} \left(\xi_k + \xi_k^* \right) \right],$$

$$\text{subject to} \quad \left(w^T \phi(x_k) - b - y_k \right) + \xi_k \geq \varepsilon,$$

$$\left(w^T \phi(x_k) - b - y_k \right) - \xi_k \leq -\varepsilon,$$

$$\xi_k \geq 0, \xi_k^* \geq 0, \; k = 1, \ldots, n,$$

As a loss function the ε-insensitive loss function is used [3]. In the functional the first term $\frac{1}{2}\|w\|^2$ requires the function to be as flat as possible, and the second term penalizes any deviations larger than the ε for all training data. Positive constant C is a regularization parameter. This optimization formulation can be transformed into the dual problem [3], and its solution is given by

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

where the dual variables are subject to constraints $0 \leq \alpha_i, \alpha_i^* \leq C$ and the kernel function $K(\cdot,\cdot)$ is a symmetric function satisfying Mercer's conditions [3]. The sample points that appear with non-zero coefficients α_i are called support vectors. The nonlinear mapping ϕ is usually implicitly defined via a kernel function $K(x, z) = \phi(x)^T \phi(z)$.

It is well known that SVM estimation accuracy depends on a good setting of hyperparameters C, ε and the kernel parameters. The problem of optimal parameter selection is very complicated because SVM model complexity depends on all three parameters. In this paper we focus on the choice of ε and kernel parameter σ.

In the widely used kernels, RBF kernels are very popular for their universal approximations. We restrict ourselves to such a class of translation invariant kernels $K(x, z) = K(\|x - z\| / \sigma)$. The parameter σ rescales the input data and determines the kernel capacity. In our experiments we use Gaussian RBF kernel $K(x, z) = \exp(-\|x - z\|^2 / 2\sigma^2)$.

For regression model quality estimation we use mean square error $MSE = (\frac{1}{n} \sum_{i=1}^{n} (f(x_i) - r(x_i))^2)^{1/2}$.

978-1-4244-2320-0/08/$25.00 ©2008 IEEE

III. EXPERIMENTAL RESULTS

To compare different methods for SVM parameters selection we used functions $r_1(x) = sin(|x|) / |x|$, $r_2(x) = sin(0.25\,x^2)$, $r_3(x) = exp(-x^2)$. We generated training set of size 100 by an additive noise process $y_i = r(x_i) + e_i$, where the inputs x_i were uniformly sampled from the domain [-10,10] and the noise e_i had Gaussian distribution with zero mean and the dispersion of 10% of the source signal power $r(x)$. Using generated data we constructed regression model using SVM.

At first we investigated the influence of kernel parameter σ for the model function $r_1(x)$. Parameter ε was fixed and equals to 0.1. When the value of the parameter σ was small we constructed highly oscillated function shown in fig. 1. The value of *MSE* in this case was 0.059 and the number of support vectors was 45.

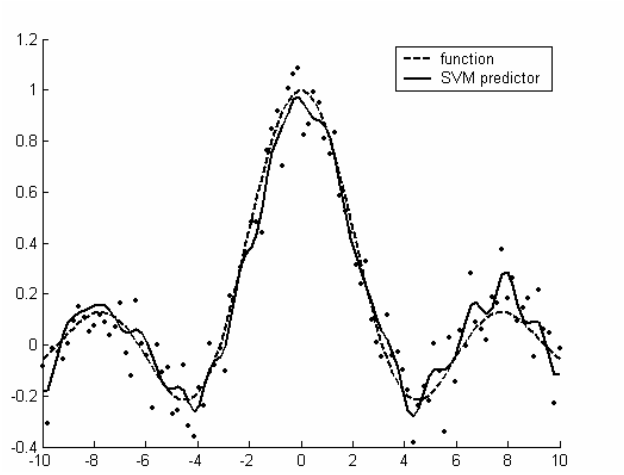

Fig. 1. SVM regression with $\sigma = 0.5$.

Using bigger value of the parameter $\sigma = 2.0$ we had the model with almost the same approximation quality but the function had better degree of smoothness. This is shown in fig. 2. The value of *MSE* in this case was 0.040 and the number of support vectors was 40. Thus we not only had lower *MSE* value but also lower regression model complexity.

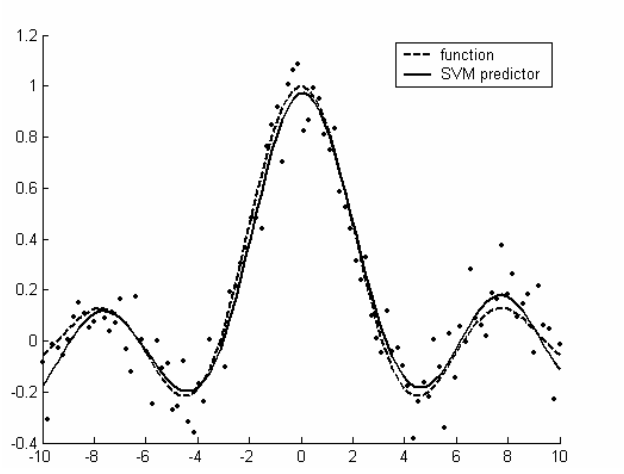

Fig. 2. SVM regression with $\sigma = 0.5$.

The influence of parameter ε is shown in fig. 3, 4. In both cases the value of the parameter σ was fixed to 2.0.

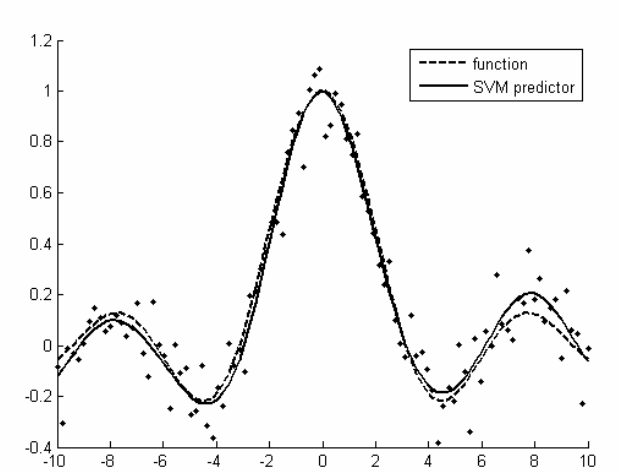

Fig. 3. SVM regression with $\varepsilon = 0.05$.

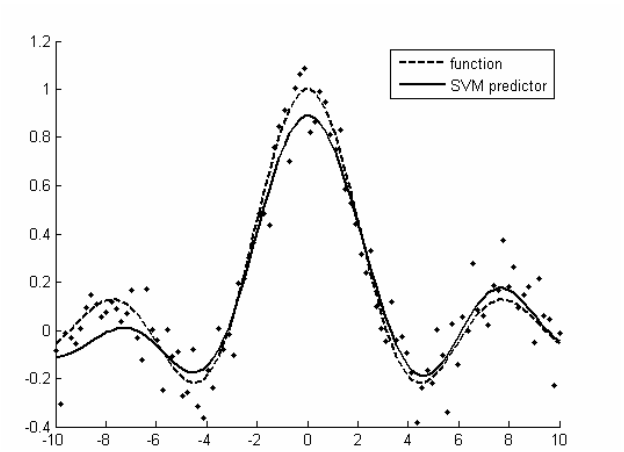

Fig. 4. SVM regression with $\varepsilon = 0.2$.

Values of MSE for both cases were 0.037, 0.064 for fig. 3 and 4 respectively. But the number of support vectors in second case decreased dramatically to 9 support vectors from 61 in the first case. Thus, when we increasing the value of ε we decrease the number of support vectors and usually increase the value of *MSE*. In general case one needs to find a compromise between model complexity and approximation quality of the model because simple model have poor descriptive properties to model even training data, but complex model have good approximation abilities on training data and bad one on testing data. So we will have the case of overtraining.

Dependence of *MSE* from the both parameters simultaneously is shown in fig. 5.

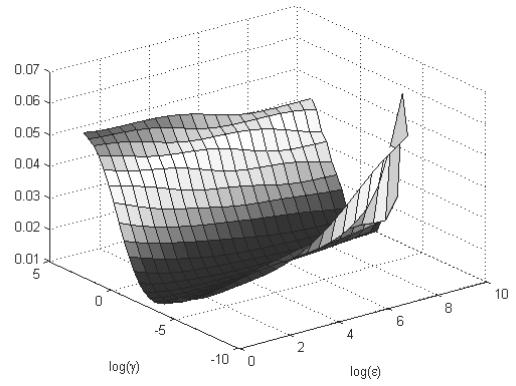

Fig. 5. SVM performance with different values of σ and ε.

Because of complex non-unimodal dependency of *MSE* from parameters σ and ε we used a grid search method to determine the optimal values of these parameters. We analyzed simple grids with different step size and nested grids. We have also used heuristic approach described in [2]. For heuristic approach parameters were set to $\sigma = 0.5 \cdot (\max(x) - \min(x)), \varepsilon = 3s\sqrt{\frac{\ln n}{n}}$, where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{y} - y_i)^2$. It is worth mentioning that this approach has no theoretical justification and does not use any prior knowledge about training data. But according to [2] heuristic approach gives good estimation of parameters in most cases.

Rages of parameters values were selected according to [4]. The range of parameter σ was set to $[10^{-3}/\rho; 1.9/\rho]$, where ρ is minimal distance between two nearest elements of training data. The range of parameter ε was between zero and the maximum distance between any of two elements of training data. Since the ranges were very wide we used logarithmical scale.

For parameters selection we used uniform grids of the sizes 5 × 5 (25 mesh points), 10 × 10 (100 mesh points), 15 × 15 (225 mesh points), 20 × 20 (400 mesh points). We have also used nested grids. For nested grids the outer grid had the size 5 × 5. On this grid was searched for the minimal value of the model quality criterion and around this value we constructed the inner grid with the size 5 × 5. The step of this inner grid was five times smaller the step of outer grid. We have also used situational optimization by each of two parameters and construction of the grid around heuristically selected parameters.

In practice, we do not know the function $r(x)$, so we can not calculate the value of *MSE*. For estimation of regression model quality we used 10-fold cross-validation criterion. Training data were spitted randomly for 10 subsets of equal size, each subset was sequentially discarded and trains the classifier 10 times, each time leaving out one of the subsets from training, but using the omitted subset to compute the prediction errors. For overall performance we calculated *MSE* value to see the real quality of regression model. Experimental results for different types of the grids are shown in Table I. In this table the number of cross-validations *CV* is a characteristic of computational complexity, *EMSE* is the value of cross-validation criterion.

TABLE I.
EXPERIMENTAL RESULTS FOR DIFFERENT PARAMETERS SEARCH METHODS

Method	*CV*	Function	*EMSE*	*MSE*
Grid 5 × 5	25	$r_1(x)$	0.01164377	0.09647524
		$r_2(x)$	0.01468316	0.08950651
		$r_3(x)$	0.01227716	0.09042292
Grid 10 × 10	100	$r_1(x)$	0.01129492	0.09548682
		$r_2(x)$	0.01361366	0.09193546
		$r_3(x)$	0.01165975	0.08769644
Grid 15 × 15	225	$r_1(x)$	0.01120984	0.09550864
		$r_2(x)$	0.01327032	0.09118706
		$r_3(x)$	0.01168978	0.08677882
Grid 20 × 20	400	$r_1(x)$	0.01119106	0.09563895
		$r_2(x)$	0.01317500	0.09094200
		$r_3(x)$	0.01165251	0.08678891
Nested grid	50	$r_1(x)$	0.01141012	0.09558591
		$r_2(x)$	0.01310579	0.09081617
		$r_3(x)$	0.01193152	0.08577250
Heuristic approach	–	$r_1(x)$	0.02504965	0.09469944
		$r_2(x)$	0.06647435	0.49331670
		$r_3(x)$	0.01766664	0.10328257
Sequential optimization	40	$r_1(x)$	0.01142714	0.09613542
		$r_2(x)$	0.01387672	0.09087674
		$r_3(x)$	0.01216794	0.08956741
Grid 5 × 5 around heuristically selected parameters	25	$r_1(x)$	0.01125791	0.09518566
		$r_2(x)$	0.06765008	0.21862768
		$r_3(x)$	0.01198660	0.08849944

Based on these experimental results we can conclude that for examined examples decrease of grid step does not substantially increase model quality. Thus regarding computational complexity it could be recommended to use bigger step size of the grid while searching for optimal parameters. It is worth mentioning that when we use heuristically selected parameters we construct smooth functions but not always they have good approximation abilities. Using these experimental results we suggest to use nested grids and to start constructing grid around heuristically selected parameters.

IV. SUMMARY

In this article we examined different methods for parameter selection using grids. Also we compared grid approach with heuristic approach. Experimental results showed that nested grids approach have advantages in computational complexity. We have also suggested the usage of heuristic approach for selecting starting values for nested grids.

REFERENCES

[1] A. Smola. *Regression Estimation with Support Vector Learning Machines. Master's thesis.* Technische Universitat Munchen, 1996

[2] V. Cherkassky and Y.Q. Ma, *"Practical selection of SVM parameters and noise estimation for SVM regression",* Neural Networks, no. 17, 2004, pp. 113–126.

[3] V. Vapnik, *The Nature of Statistical Learning Theory.* Springer, N.Y., 1995.

[4] C.M. Huang, Y.J. Lee, *"Model selection for support vector machines via uniform design",* Computational Statistics & Data Analysis, no. 52, 2007, pp. 335–346.