The Method of Applying Support Vector Machine to Engineering Data Regression

Jin Tian

Information College, Capital University of Economics and Business, Beijing, China

tian@cueb.edu.cn

Keywords: support vector regression; machine learning; kernel function; parameter optimization

Abstract. Based on machine learning concepts, this paper has put forward two key problems of the application of support vector regression and has given a solution to these problems. It is that the different characteristics of the sample data are decisive factors of the schemes of the selection, and the procedure of the structure of kernel function and parameter optimization are proposed.

Introduction

The data-based machine learning makes use of the limited number of training samples to establish a dependency of input and output. Its mathematical form is as follows. Let $X \subseteq R^m$ be the pattern space, $Y \subseteq R^n$ be the output space, $Z=X \times Y$ be the sample space, $x \in X$ be a pattern, $y \in Y$ be an output, $z=(x,y)$ ordered pair forms a sample. If there is a fixed but unknown dependency $F(x,y)$ between X and Y in Z (also known as the joint probability distribution), and it is given the sample set $Z_0=\{(x_i,y_i)\}_{i=1}^{l}$ which contains l independent identical distribution samples according to $F(x,y)$, then the aim of machine learning is to select the optimal function $f(x,\omega^*)$ in a specified set of functions $\aleph=\{f(x,\omega)|\omega$ is a set of parameters$\}$ to approximate function $F(x,y)$. Generally, measuring the degree of approximation uses the actual risk function

$$R(f)=\int L(y,f(x,\omega))\, dF(x,y). \tag{1}$$

It is the average of loss function $L(y,f(x,\omega))$ in the sample space Z. It is also called the approximation error of the generalization error. The function approximation problem may select the loss function $L(y,f(x,\omega))$ as follows.

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

or

$$L(y,f(x,\omega))=|y- f(x,\omega)|. \tag{3}$$

Above all, the equivalent condition for credible prediction is $R(f)$ minimized. Because $R(f)$ depends on $L(y,f(x,\omega))$, and according to the loss function $L(y,f(x,\omega))$ of the regression, we can qualitatively extrapolate that the estimated $R(f)$ is closely related to the training set $Z_0=\{(x_i,y_i)\}_{i=1}^{l}$ and the specified function class $f(x,\omega)$. Therefore, this paper proposes two key problems in machine learning: the selection of forecast factors and the quality of sample set is the prerequisite foundation, and it is the key core that selecting the optimal function $f(x,\omega^*)$ in a specified function class $f(x,\omega)$ make $R(f)$ the smallest.

The Strategy of Problem Solving

The Construction of the Prerequisite Foundation

The prerequisite foundation refers to the attributes of the samples, sample quality, sample data pre-processing. The observation sample must reflect the dependency relationship between X and Y. Therefore, the attributes of the input x and the output y is determined according to the specific application of business domain knowledge. To select the attributes of the input x, two basic principles should be followed:

First, choose the factor which has great influence on the output and contains sufficient information of the output.

Secondly, there is irrelevant or very little correlation between the various attributes. If it cannot be determined whether an attribute can be used as the input or nor, the machine learning sample should be established separately with or without this attribute. Whether the attribute is selected or not should be determined by the quantity of the generalization error.

Sample quality refers to the size of the sample. It should reflect the diversity, representativeness, and the accuracy of the data collection. The sample set Z_0 should be spread over the sample space Z, in order to fully contain unknown dependencies. The "interpolation" could obtain more reliable results, but the "extrapolation" is often difficult to guarantee reliability. So it is important to avoid extrapolation by taking advantage of the diversity of the representative samples. Practices have shown that the more complex the input-output relationship are and the greater the noise in the sample is, the more samples are required to guarantee the accuracy of mapping, the more accurate the training results reflect its inherent laws. For the samples are obtained through experiments, according to the concrete business content, the sample size is determined in accordance with the test design requirements.

Pre-processing of sample data is to analyze the data with different dimension and to map the raw data to a particular numerical range by some function transform. By this process, it can guarantee that attributes with larger value range have weights equal to those with smaller value range. Furthermore, it can reduce the prediction error. Function transform can be categorized into several types: center transformation, logarithmic transformation, normalization transformation, auto scaling treatment, standardized treatment. To select appropriate type of transformation, features of the sample data and transform characteristics should be evaluated. If any component of input-output is not of the same dimension to the others, that component is to be transformed within its own range. If there is a correlation between the attributes, the correlated attributes can be simplified as irrelevant virtual features by principal component analysis and kernel principal component analysis.

The Key Core Determination

The key core aims to select the optimal function $f(x,\omega^*)$ in a specified function set $f(x,\omega)$ to make $R(f)$ the smallest. Based on the probability law of large numbers, previous machine learning methods are to make sample empirical risk

$$R_{emp}(f) = \frac{1}{l}\sum_{i=1}^{l} L\left(y_i, f(x_i,\omega)\right). \tag{4}$$

approximate the actual risk function, to replace the minimum of actual experience risk function $R(f)$ with the minimum empirical risk $R_{emp}(f)$ of parameter ω. However, there is no reliable theoretical basis for this. The large number theorem only shows that $R_{emp}(f)$ tends to $R(f)$ in the sense of probability when the number of samples tends to infinity, but does not guarantee that ω^* making $R_{emp}(f)$ minimum and $\omega^{\sim}$ making $R(f)$ minimum are the same, and furthermore, does not guarantee that $R_{emp}(f(x,\omega^*))$ will be able to approach $R(f(x,\omega))$.Even if there are these two conditions, the quality of the actual training set is limited and could not be infinite. Therefore, the traditional machine learning cannot ensure that the actual risk converges to a minimum. Hence, machine learning methods should solve the following four problems:

①How to determine the alternative function set $f(x,\omega)$?

②Besides the empirical risk minimization principle, on what basis the optimal function $f(x,\omega^*)$ is determined?

③Learning models based on the different capacity of sample are not exactly the same. Do they tend to converge with the increase of the sample quantity? How good are their convergence speeds?

④How to evaluate the generalization performance of the learning model?

Vapnik[1] specializes in the machine learning of small sample quantity, and established statistical learning theory. Under the direction of this theory, support vector machine (SVM) is a practical tool

established based on precise mathematical reasoning. Issue②③④mentioned before have been effectively resolved in the SVM. Issue① will be made up perfectly with the rapid development of the kernel methods.

The SVM method is compared with other machine learning methods by Meyer D[2]. With regard to classification, the SVM is overwhelming compared to other 16 kinds of classification methods including neural network, classification tree, nearest neighbor method in accuracy, generalization ability and modelling computation. With respect to regression, the SVM has obvious advantages over other 9 kinds of regression, such as neural network. In addition, the SVM has prominent advantages for specific data set. Therefore, this paper takes SVR as the solution of "the key core" in function approximation. The remainder of this paper is to explore the selection of ε-SVR application model.

The Specific Methods and Measures

The Selection and Construction of Kernel Function. Relatively good selection and construction methods of kernel function are as follows.

The Selection of RBF Kernel Function. Research shows that selecting RBF kernel function is better than selecting other kernel function when the prior knowledge are lacking. RBF kernel function is a universal kernel function. It can obtain better generalization performance without any prior knowledge, and it is suitable for any distribution sample through choosing of the parameters [3,4].

The Selection of a Combination of Kernel Functions. When samples contain heterogeneous information, are large scale, or data distribution is uneven in a high dimensional feature space, it is not reasonable that using a single kernel function map all samples. Theory and application has proved that it can enhance the interpretability of the decision function for multiple single kernels together. Common form is a convex combination of multiple basic kernels:

$$K = \sum_{i=1}^{m} \beta_i K_i, \beta_i \geq 0, \sum_{i=1}^{m} \beta_i = 1 \quad (5)$$

Smits studies mapping features of representative global kernel (polynomial kernel) and local kernel function (RBF kernel), and establish a synthetic kernel which has a better learning ability and a better extrapolation capability:

$$K - \rho \cdot \exp^{-\|x - x_i\|/2\sigma^2} + (1 - \rho)(x \cdot x_i + 1)^d \quad (6)$$

The Construction of Multi-scale Kernel Functions. The selection of basic kernels with suitable parameters for combined kernel is the lack of scientific guidance. And further, combined kernel does not satisfactorily resolve uneven distribution of the sample, and limits the representation ability of decision function. Referencing the scale space theory in the research of computer vision, the kernel functions with multi-scale representation is combined into a multi-scale kernel. This multi-scale kernel has the more flexible and more complete ability of multi-scale selection than the combined kernel, and is based on more scientific theory. Zheng [5,6] and Yang[7] proposed multi-scale SVR. They used respectively for non-flat function estimation and time-series forecasting. Two kinds of multi-scale kernel function can be constructed by using the RBF kernel and the wavelet kernel function.

Parameter Optimization. In essence, the parameter selection of SVR is the basal optimization. This corresponds with choosing an appropriate feature space. The training of SVR is the high-level optimization. This correspond with solving the optimal hyper plane for a given feature space. So the parameter optimization needs repeated iteration, to minimize generalization errors of the corresponding kernel function.

The Meaning and Effects of the Parameters. The ε-SVR on the RBF kernel need to determine three parameters: insensitive parameter ε, penalty coefficient C, width parameter σ. The optimal regression function of the ε-SVR is "the flattest" function that meets ε pipeline restriction in the specified set.

Parameter ε defines the insensitive band width. Its value is closely related with the sample noise. The smaller ε value is, the more support vectors there are, the more tortuous optimal regression function is, and the higher regression precision there is, the better fitting there is. But, if the ε is too small, regression efficiency decreases. The bigger ε value is, the fewer support vectors there are, and the greater regression error is. If ε is too big, the model will become simple. This will result in reducing the accuracy of regression. Correlation research shows that the generalization error is not sensitive to Parameter ε. We should take ε∈[0,0.2] according to literature [8].

The parameter C is the punishment degree of the samples outside ε insensitive band, and also is the ratio of confidence interval and empirical risk. It makes learning machine's extend ability to the best. If C is too small, the punishment to samples outside ε insensitive band is small, and the training error will become large. This leads not enough learning, and the lowering of model generalization ability. If the C value is too large, the optimal hyper plane weight ||ε|| gets smaller. This makes support vectors increase, and the training error become small. It leads to exceed learning, and to enlarge the generalization error. For each training set there is an optimal C of better generalization performance. In addition, Lagrange multipliers which correspond to the support vectors are associated with C. When C is too large, it will make the corresponding support vectors plays a decisive role. When C is small, Lagrange multipliers are relatively close; the corresponding support vectors' contributions to the regression model are much the same. So, choosing the appropriate C, ε can make the regression model is not sensitive to outliers. C∈[1,108] is taken in literature [8].

The width parameter σ of RBF kernel implicitly determines the mapping function and the feature space. This parameter has an important influence on the performance of SVR. If σ is too large, it will lead to regression function gentle. If σ is too small, it results in over-fitting and brings the memory effect. The influence of parameter σ on the generalization ability is similar to scale parameter. In a stable range of parameters σ, there is a theoretical basis for the fact that Generalization ability is stronger. So just find a parameter value within this range.

Parameter Optimization Methods. The generalization ability of SVR depends on ε, C, σ three parameters. Selecting them is an optimization problem. Due to the coupling effect of various parameters on the generalization, it is not reasonable that each parameter is optimized separately. The parameters ε, C, σ should be optimized at the same time. The available are cross-validation, experience selection, a trial and error methods, gradient descent methods, Bayesian methods etc. In addition, particle swarm optimization, genetic algorithms and other intelligent optimization method is also used in the SVM parameter optimization.

The paper thinks that quantum particle swarm optimization (QPSO) is an ideal choice for SVR parameter optimization. This is not only that QPSO inherited the advantages of the PSO with simple algorithm, fewer parameters and search quickly, etc., but also on the following advantages [9]:

It is a global convergence algorithm that can guarantee to convergence to the global optimal solution; the algorithm needs only one control parameter; it is more convenient for application; standard test functions show that QPSO has better solving performance.

There have been many improvement measures in controlling QPSO premature convergence. There is a great probability that high-precision global optimal solution is found.

For SVR parameter Optimization, this paper takes the following four points:

Ÿ In order to accurately calculate the fitness function, particles are not directly formed with ε, C, σ parameters. The particles are formed with $10\varepsilon,\ \log_{10} C, \log_{10} 1/2\sigma^2$. So, these parameter values are more reasonable in the same order of magnitude, they are converted to the range [0,2], [-2,8], [-3,5] as in literature [8].

Ÿ Considering the sample image distribution is very likely not flat, in order to eliminate as much as possible the influence of the sample away from the regression model, the penalty parameter C is weighted processing as in literature [10].

Ÿ According to the literature [11], the multi-scale RBF kernel is formed. According to the literature [10], the multi-scale wavelet kernel is formed.

Ÿ A hybrid algorithm is composed of QPSO and leapfrog algorithm to speeding up the parameter optimization and shortening the training time of SVR.

Conclusion

For application of SVR, it is necessary to determine samples framework, kernel selection and parameter optimization. About kernel, we may choose especially RBF universal kernel, the multi-scale RBF kernel and the multi-scale wavelet kernel in order to improve generalization performance. By QPSO, parameter optimization will be speed to shorten the training time of SVR.

Acknowledgement

The research work was supported by Beijing Natural Science Foundation under Grant No. 20873999.

References

[1] Vapnik V N., Statistical Learning Theory, Trans. Xu Jianhua, Zhang Xuegong, Beijing: Electronic Industry Press, Beijing 2004.

[2] Meyer D, Leisch F, Hornik K., Benchmarking Support Vector Machines, Report No 78. Vienna University of Economics and Business Administration, Austria, 2002.

[3] Zien A, Ratsch C C, Mike S., Engineering support vector machines kernels that recognize translation initiation sites in DNA, Bioinformatics, 16, pp.799-807, 2000.

[4] Wu Tao, Kernel function properties, methods, and its application in fault detection, National University of Defense Technology, PhD thesis, Beijing, 2003.

[5] Zheng D N, Wang J X, Zhao Y N, Non flat function estimation with a multi-scale support vector regressions, Neurocomputing, vol.70, pp.420-429, 2006.

[6] Zheng D N, Wang J X, Zhao Y N, Time series predictions using multi-scale support vector regressions, Lectur Notes in Computer Science, vol.3959, pp.474-481, 2006.

[7] Yang Z, Gao J, Xu W, Multi-scale Support Vector Machine for regression, ISNN'06, Incs 3971, pp.1030-1037, 2006.

[8] B Ustun, W J Melssen, Determination of optimal support vector regression parameters by genetic algorithms and simplex optimization, Analytical Chimica Acta(S0003-2670), vol.544(1-2), pp.292-305,2005.

[9] Sun Jun, Fang Wei, Wu Xiaojun, Xu Wenbo, Quantum-behaved Particle Swarm Optimization: Theory and Application, Beijing: Tsinghua University Press, 2011.

[10] Yu Yanfang. Improved Support Vector regression, its Application in Process Modeling and Contrl. East China University of Science and Technology, PhD thesis, 2010.

[11] Wang Hongqiao, Cai yanning, Multiple Kernel Methods for Pattern Analysis and its Application, Beijing: National Defense Industry Press, 2014.