A Brief Overview on Parameter Optimization of Support Vector Machine

Qi Cao¹, Lei Yu², Mingsheng Cheng³
¹Department of Training, Logistical Engineering University, Chongqing, 401311, China
²College of Computer and Information Science, Chongqing Normal University, Chongqing, 401331, China
³Logistics College, Naval University of Engineering, TianJin, 300450, China

ABSTRACT: Support vector machine (SVM) has been successfully applied in classification and regression problems. But it is very sensitive to the selection of parameters. The fundamental principles of SVM are analyzed firstly. The main optimization methods and achievements for SVM parameters are introduced. And the popular fitness functions used for the parameter optimization of SVM are described. The objective of this paper is to provide readers a brief overview of the recent advances for parameter optimization of SVM and enable them to develop and implement new optimization strategies for SVM-related research at their disposal.

KEYWORDS: Support vector machine; Parameter selection; Optimization algorithm; Fitness function.

1 INTRODUCTION

Support vector machine (SVM) is a universal learning method developed from statistical learning theory, which improves the generalizing ability by structural risk minimization [1]. SVM can effectively solve the problems characterized by few samples, non-linearity, high dimensions and local minima. In the past 20 years, SVM has been successfully applied in classification and regression problems [2-5]. However, SVM is very sensitive to its own parameters and usual methods for choosing parameters based on exhaustive search will be intractable as soon as the number of parameters exceeds limitation [6], which has become the main obstacle to its applications.

Recently many researchers have studied the selection and improvement of SVM parameters. These researches can be mainly divided into two kinds of methods, i.e. grid search and optimization methods [7]. Grid search is an improved exhaustive searching through a manually specified subset of the parameter space of a learning algorithm. Since the parameter space of SVM may include real-valued or unbounded value spaces for certain parameters, manual setting of boundaries and discretization are necessary before applying grid search. The parameter selection based on grid search will be straightforward and simple [8], but it still meets time complexity problem and inaccurate results if there are more parameters or improper settings of parameter ranges and steps. Optimization methods change the parameter selection of SVM into an optimization problem which minimize the fitness function of SVM under certain constraint conditions and minimize the estimation values of generalization ability based on the optimal solutions [7]. Compared with grid search, the optimization methods, usually involving the soft computing algorithms, can often obtain better solutions in less time [9]. So the parameter optimization of SVM is reviewed in this paper.

In next section, the fundamental principles of SVM are given. In Section 3, the main optimization methods and achievements for SVM parameters are presented. Section 4 describes the popular fitness functions used for the parameter optimization of SVM. Finally, concluding remarks are given in section 5.

2 FUNDAMENTAL PRINCIPLES OF SVM

2.1 SVM for classification (SVC)

Given sampled data \((x_i, y_i)_{i=1}^n\), where \(x_i \in \mathbb{R}^d\) is \(n\)-dimensional sample input and \(y_i \in \{-1,+1\}\) is the class label, the classification is to obtain a corresponding label \(y\) according to \(x\). The training patterns are said to be linearly separable if a vector
\( w^T \) which determines the orientation of a discriminating plane and a scalar \( b \) which determines the offset of the discriminating plane from origin can be defined so that the following equation is satisfied [10].

\[
y_i(w^T \varphi(x_i) + b) \geq 1, \quad i = 1, 2, \cdots, n
\]  

(1)

where the non-linear function \( \varphi(*) \), \( R^n \to R^b \) maps input space into a high-dimensional feature space. Based on structural risk minimization theory [1], the classification is equal to solving the following optimization problem:

\[
\text{minimize}_{\alpha, \xi} J = \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} \xi_i
\]

subject to \( y_i(w^T \varphi(x_i) + b) \leq 1 - \xi_i, \quad i = 1, 2, \cdots, n \)

where constant \( C \) is the penalty parameter and \( \xi \) are the slack variables. The proportion of confidence intervals and empirical risk is adjusted by \( C \). By using Lagrangian method, the following quadratic programming problem is obtained:

\[
\text{maximize}_{\alpha} J = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{n} \alpha_i
\]

subject to \( \sum_{i=1}^{n} \alpha_i y_i = 0 \) 

\( 0 \leq \alpha_i \leq C, \quad i = 1, 2, \cdots, n \) 

Based on the Mercer condition, the kernel methods only deal with the kernel function, which is shown as below:

\[
K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)
\]

(4)

Finally, in dual space the nonlinear SVM is replaced by the following:

\[
y(x) = \text{sign} \left( \sum_{i=1}^{n} \alpha_i y_i K(x_i, x_j) + b \right)
\]

(5)

The non-zero Lagrange multipliers \( \alpha_i \) are called support values. The bias \( b \) can be computed using the Karush-Kuhn-Tucker (KKT) conditions.

2.2 SVM for regression (SVR)

The function approximation problem is finding a function \( f(x) \) that can obtain a corresponding \( y \) according to \( x \), which does not belong to the sample after sample-based training. This function is hypothesized as follows [10]:

\[
f(x) = w^T \varphi(x_i) + b \quad w \in R^b, b \in R
\]

(6)

The solution is finding \( w \) and \( b \) to make \( |f(x) - w^T \varphi(x_i) - b| < \varepsilon \) according to \( x \), which can be solved by the optimization problem shown as follows:

\[
\text{minimize}_{w, b, \varepsilon} J = \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)
\]

subject to \( y_i - w^T \varphi(x_i) - b \leq \varepsilon + \xi_i \) 

\( w^T \varphi(x_i) + b - y_i \leq \varepsilon + \xi_i^* \) 

\( \xi_i, \xi_i^* \geq 0 \)

(7)

where \( C \) shows the tradeoff between the smoothness of \( f(x) \) and error allowance; \( \xi \) and \( \xi^* \) are slack variables; \( \varepsilon \) is error probability which is shown as follows:

\[
|y - f(x)| \leq \varepsilon, \quad |y - f(x)| - \varepsilon, \quad \text{others}
\]

(8)

The Lagrange method and kernel functions are used to solve this quadratic programming problem which is shown as below:

\[
w = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \varphi(x_i)
\]

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

(9)

2.3 Main kernel functions

Main kernel functions in machine learning theory include the following:

- linear kernel function: \( K(x_i, x_j) = x_i \cdot x_j \); 
- polynomial kernel function: \( K(x_i, x_j) = (\gamma(x_i \cdot x_j) + r)^d \), \( \gamma > 0 \); 
- radial basis function kernel function: \( K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2) \), \( \gamma > 0 \); 
- sigmoid kernel function: \( K(x_i, x_j) = \tanh(\gamma(x_i \cdot x_j) + r) \); 

where \( \gamma \), \( r \), and \( d \) are kernel function parameters. 

In general, the SVM parameters mainly include penalty parameter \( C \), error probability parameter \( \varepsilon \) and kernel function parameters, which are the objects of parameter optimization of SVM.

3 MAIN METHODS FOR PARAMETER OPTIMIZATION OF SVM

3.1 Genetic algorithm (GA)

operators (IO-GA) to optimize the SVM classifier’s parameters. Qin et al. [15] formulated parameter optimization of SVM for classification based on niche genetic algorithm (NGA) which avoided the premature situation and better maintained the diversity of solution.

3.2 Particle swarm optimization (PSO)

PSO is an adaptive algorithm based on a social-psychological metaphor: a population of individuals adapts by returning stochastically toward previously successful regions. Lu et al. [16] presented a new SVM model based on PSO algorithm for parameters optimization and applied this model in short-term load forecasting of electric power system. Wang et al. [17] adopted a chaotic PSO algorithm modified by virtue of chaotic motion with sensitive dependence on initial conditions and ergodicity to select parameters of SVM. Yao et al. [18] used comprehensive learning PSO (CLPSO) to optimize the kernel parameters of the SVM for remote sensing image classification. Miranda et al. [19] proposed a prototype in which multi-objective PSO (MOPSO) and crowding distance mechanism (CDR) was used to select the values of two SVM parameters for classification problems. İlhan et al. [20] optimized $C$ and $\gamma$ parameters of SVM by using PSO algorithm to predict single nucleotide polymorphisms (SNPs).

3.3 Differential evolution (DE)

DE is a popular evolutionary algorithm and primarily suited for numerical optimization problems, which is well-known for its simple and easy-to-understand concept, high convergence characteristics and robustness. Shu et al. [21] employed a DE-SVM model that hybridized the DE and SVM to improve the classification accuracy for rainstorm forecasting. Kryš et al. [22] applied DE to tune the hyperparameters of least-square SVM classifier on a signal-averaged electrocardiography dataset. Fu et al. [23] utilized a new SVM-DE model that hybridized the DE technique and SVM to resolve the problem of force identification. Bhadra et al. [24] devised a meta classifier model by using SVM and optimized its kernel parameters by using DE. Li et al. [25] proposed an SVM classification system based on DE to improve the generalization performance of the SVM classifier.

3.4 Other optimization algorithms

There are other algorithms to be used to optimize the SVM parameters. Lin et al. [26] proposed a simulated annealing (SA) based approach to optimize the parameter values for SVM and obtain a subset of beneficial features. Zhang et al. [27] adopted ant colony optimization (ACO) approach to present a novel ACO-SVM model for parameter optimization problem of SVM. Aydin et al. [28] presented a multi-objective artificial immune system (AIS) to optimize the kernel and penalize parameters of SVM and applied it to fault diagnosis of induction motors and anomaly detection problems. Bai et al. [29] proposed parallel artificial fish swarm algorithm (PAFSA) to optimize kernel parameter and penalty factor of SVM and applied the optimal parameters to a speech recognition system. Ao et al. [30] proposed a SVM parameter optimization method based on artificial chemical reaction optimization algorithm (ACROA) to diagnose roller bearing faults. Afif et al. [31] used a scatter search approach to find near optimal values of the SVM parameters. Kartelj et al. [32] introduced an electromagnetism-like (EM) approach for solving the problem of parameter tuning in the SVM.

In addition, the combination methods become popular gradually, which are complementary to different optimization algorithms. For example, Shi et al. [33] presented the combined algorithm based on quantum-behavior PSO (QPSO) and SA to optimize the parameters of SVM; Li et al. [34] proposed a novel hybrid optimization algorithm based on immunity clone (IC) and DE for parameter selection of SVM.

4 FITNESS FUNCTIONS FOR PARAMETER OPTIMIZATION OF SVM

Fitness function is an important factor for evaluation and evolution of SVM providing satisfactory and stable results in real-world applications. The simplest fitness function for parameter optimization of SVM is the accuracy of SVM [28, 30] or the number of support vectors [28]. Here are the other popular fitness functions which can be used in different situations.

i) Seo [12] used SVM classification accuracy and the number of selected features to design a fitness function. The fitness function is designed as follows:

$$F_{fit} = W_1 \times Accuracy + W_2 \times Nonzeros$$

where $W_1$ is the weight for SVM classification accuracy, $Accuracy$ is SVM classification accuracy, $W_2$ is the weight for the number of features and $Nonzeros$ is the number of selected features.

ii) Qin et al. [15] evaluated the classification performance of SVM with new leave-one-out known as NLOO. The fitness function is shown as below:

$$F_{fit} = \frac{1}{\sum_{j=1}^{d} \sum_{(x_i, y_i \in D_j)} (y_i - f(x_i))^2 + \alpha}$$
where all samples are divided into \( d \) groups and every group is represented by \( D_j, j = 1, 2, ..., d \); consider one group as testing samples and the remainder \( d-1 \) groups as the training samples; the evaluation function of the \( j \)-th sample is obtained by the reciprocal of error sum squares about the result output of \( d \) groups with SVM classifier and objective function.

iii) Lu et al. [16] defined the fitness function as below:

\[
F_{fit} = \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{1}{m} \sum_{k=1}^{m} (t_{jk} - y_{jk})^2 \right]^{1/2} 
\]  

(12)

where \( t_{jk} \) is the target value, \( y_{jk} \) is the output value, \( m \) is the nodes of output, \( n \) is the number of training data.

iv) Wang et al. [17] used \( k \)-fold cross validation error as fitness function which is calculated as below:

\[
F_{fit} = \frac{1}{k} \sum_{j=1}^{k} \left[ \frac{1}{l_j} \sum_{i=1}^{l_j} (y'(x_i) - y(x_i))^2 \right]^{-1/2} 
\]  

(13)

where \( y'(x_i) \) is the forecast value for output of SVM, and \( y(x_i) \) is target output value, and \( l_j \) denote the number of \( j \)-th sample subset, \( j = 1, 2, ..., k \).

v) Fu et al. [23] evaluated the fitness function of each individual by the follows:

\[
F_{fit} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (y_{pre}(i) - y_{ori}(i))^2 
\]  

(14)

where \( n \) is the number of sample points, \( y_{pre} \) and \( y_{ori} \) are predicted and original values, respectively.

vi) Zhang et al. [27] put forward the fitness function as below:

\[
F_{fit} = \prod_{i=1}^{l} \psi (-y_i f(x_i)) 
\]  

(15)

where \( l \) is the number of validation set samples; \( \psi \) is a step function: \( \psi(x) = 1 \), when \( x > 0 \), otherwise \( \psi(x) = 0 ; f \) is the decision function of SVM.

vii) Aydin et al. [28] also described two fitness functions shown as follows:

\[
F_{fit} = \frac{nsv}{m} 
\]  

(16)

where \( m \) is number of training data and \( nsv \) is number of support vectors.

\[
F_{fit} = \frac{1}{2} \left( \frac{m^-}{m} + \frac{\pi}{m} \right) 
\]  

(17)

where \( m^- \) denotes the number of negative class records in the data set and \( \pi \) is the number of negative class records that have been classified correctly with similar meanings for \( \pi^+ \) and \( m^+ \).

5 CONCLUSIONS

SVM is very sensitive to its own parameters which decide the classification or prediction performance directly. Compared with grid search, the optimization methods can often obtain better solutions in less time. This paper presents a brief overview on the parameter optimization of SVM. Based on the analysis of SVM fundamental principles, the main optimization methods and popular fitness functions for SVM parameters are introduced. Future research may focus on the application of combined optimization methods and the selection of appropriate fitness functions for parameter optimization of SVM.

6 ACKNOWLEDGEMENTS

The research work was supported by the State Scholarship Fund of China under Grant No. 201203170027 and the National High Technology Research and Development Program of China (863 Program) under Grant No. 2012AA063501.

REFERENCES


