Measurement 58 (2014) 147–153

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Super-parameter selection for Gaussian-Kernel SVM based on outlier-resisting

Xuesong Wang *, Fei Huang, Yuhu Cheng

School of Information and Electrical Engineering, China University of Mining and Technology, 221116 Xuzhou, China

ARTICLE INFO

Article history:
Received 10 October 2013
Received in revised form 11 June 2014
Accepted 14 August 2014
Available online 3 September 2014

Keywords:
Support vector machine
Super-parameter selection
Outlier-resisting
Classification accuracy
Computational complexity

ABSTRACT

The learning ability and generalizing performance of the support vector machine (SVM) mainly relies on the reasonable selection of super-parameters. When the scale of the training sample set is large and the parameter space is huge, the existing popular super-parameter selection methods are impractical due to high computational complexity. In this paper, a novel super-parameter selection method for SVM with a Gaussian kernel is proposed, which can be divided into the following two stages. The first one is choosing the kernel parameter to ensure a sufficiently large number of potential support vectors retained in the training sample set. The second one is screening out outliers from the training sample set by assigning a special value to the penalty factor, and training out the optimal penalty factor from the remained training sample set without outliers. The whole process of super-parameter selection only needs two train-validate cycles. Therefore, the computational complexity of our method is low. The comparative experimental results concerning 8 benchmark datasets show that our method possesses high classification accuracy and desirable training time.

1. Introduction

Support vector machine (SVM) is a kernel learning method established by Vapnik on the basis of statistical learning theory [1]. SVM can well resolve the small-sample, nonlinear and high-dimensional problem. As its training is a problem of convex quadratic programming, it can be ensured that the extremal solution found is just the global optimal solution. With excellent learning ability and generalizing performance, SVM is widely applied in such fields as pattern recognition [2], regressive modeling [3], and signal processing [4]. However, to a great extent, the performance of SVM depends on the reasonable selection of super-parameters in practice. Improper super-parameters will cause the underfitting or overfitting of SVM, not only affecting the classification accuracy, but also resulting in excessively long training time and even the failure of obtaining the optimization model in limited steps [5]. Therefore, selecting proper super-parameters is an urgent problem to be resolved in the application of SVM.

The conventional super-parameter selection methods include empirical selection, grid search, gradient descent, and so on.

(1) The empirical selection method: this method requires for relevant background knowledge of issues under study, and it is extremely difficult to obtain proper super-parameters in practice.
(2) The grid search method: this method ignores the analysis on the influence of super-parameters on the generalization performance of SVM, causing the waste in terms of most computation on the training of valueless super-parameter area.

* Corresponding author. Tel.: +86 516 83590868; fax: +86 516 83590817.
E-mail address: wangxuesongcumt@163.com (X. Wang).

http://dx.doi.org/10.1016/j.measurement.2014.08.019
0263-2241/

(3) The gradient descent method: by taking into account the relationship between super-parameters and generalization performance, the gradient descent method can greatly reduce the computational complexity, but the initial point for optimization has a major influence on the super-parameters finally selected. Therefore, this method easily falls into the local optimum situation.
(4) Other methods: In recent years, with the inspiration of intelligent optimization methods, such as genetic algorithm (GA) and particle swarm optimization (PSO), succeeding in such areas as process control, economic forecasting and engineering optimization, some scholars have successively studied how to optimally select the super-parameters of SVM by applying intelligent optimization methods [6,7]. However, the optimizing performance of GA and PSO also depends on the reasonable selection of their several running parameters to a great extent. Furthermore, these methods are also observed with such problems as high computational complexity and difficulty in selecting super-parameters for large-scale dataset. Keerthi et al. [8] proposed a selection method of two-step linear search for super-parameters. The simulation result indicates that its accuracy is close to that of the grid search method, but its computational complexity is obviously reduced. Varewyck and Martens [9] define the kernel parameter by establishing the Gauss distribution model on the one hand; on the other hand, they select a relatively ideal initial value for the penalty factor by analyzing the distribution properties on different datasets, and then determining the optimal penalty factor by using the gradient search method.

Except for empirical selection, all the aforementioned super-parameter selection methods view the selection of super-parameters as an optimization problem. In another word, they all need to design proper evaluation indexes for super-parameter performance to appraise the selected super-parameters. The earliest relevant research was based on the principle of structural risk minimization. This principle grounds the theory of SVM. It states that the classification error is bounded by the sum of the training error and a supplementary confidence term written as a function of the Vapnik–Chervonenkis (VC) dimension [1]. However, in the actual situation, it is extremely difficult to estimate the exact VC dimension and thus the upper bound of classification error has no strict close association with constraints. The common evaluation index for super-parameter performance is the mean error obtained through K-fold cross-validation (KF-CV) on the training sample set [10]. The KF-CV requires one or more train-validate cycles to be performed. Each such cycle consists of one or more trials in which a SVM is trained on a training sample set, and evaluated on a validation set. If K is 5, the 5F-CV means that 5 trials per train-validate cycle are required. This method is observed with the potential risk of high computational complexity, non-rigorous index, introduction of new variables and easy falling into local optimum. The relevant approximation methods substituting KF-CV include the Xi-alpha error bound and the generalized approximate cross-validation measure. Duan et al. [11] pointed out that the computational complexity of such evaluation indexes for super-parameter performance is greatly lowered, but the obtained generalization performance of SVM is not superior to the result of 5F-CV.

The computational complexities of the foregoing super-parameter selection methods are relative high. Therefore, for the large-scale training sample set or huge parameter space, they are impractical. In addition, although the currently popular methods based on the optimization theory can obtain super-parameters with desirable generalization performance, when the spatial distribution of samples is observed with outliers, optimizing strategies often lead to the failure of super-parameter optimization process to converge quickly due to the disturbance of outliers. On the basis of analyzing the relationship between outliers in training sample set and generalization performance of SVM, we transform the super-parameter selection for SVM into the outlier-resisting. Upon analyzing, outliers can be detected by assigning a special value to the penalty factor. To resist the influence of outliers on generalization performance of SVM, outliers are excluded from the training sample set, and then the remaining training sample set is used to obtain the optimal penalty factor. In the whole super-parameter selection process, only two train-validate cycles are required to obtain proper parameters, without the huge redundant computation using the optimization-based methods.

2. Description of SVM

SVM maximizes the classification margin by constructing an optimal decision hyperplane. For the issue with supervised classification, set the training sample set $\Omega = \{(\boldsymbol{x}_i, y_i) | \boldsymbol{x}_i \in R^k\}$, $y_i \in \{-1, 1\}$, $i \in 1, 2, \cdots, l$, where, $\boldsymbol{x}_i$ represents the training sample in a k-dimensional feature space and y_i represents positive or negative class label corresponding to $\boldsymbol{x}_i$. The decision hyperplane is defined by:

$$\boldsymbol{w}\boldsymbol{x} + b = 0, \tag{1}$$

where $\boldsymbol{w}$ represents weight vector and b represents bias. Therefore,

$$y_i(\boldsymbol{w}\boldsymbol{x}_i + b) \geqslant 1. \tag{2}$$

The decision function is:

$$f(\boldsymbol{x}) = \operatorname{sgn}(\boldsymbol{w}\boldsymbol{x}_i + b), \tag{3}$$

where $\operatorname{sgn}(\cdot)$ is the following sign function:

$$\operatorname{sgn}(a) = \begin{cases} 1, & a \geqslant 0 \\ -1, & a < 0 \end{cases}. \tag{4}$$

The problem of obtaining optimal hyperplane can be transformed into the following primal quadratic programming problem [1]:

$$\boldsymbol{w}, \hat{b}, \hat{\xi} = \arg\min_{\boldsymbol{w}, b, \xi} \left(\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{l} \xi_i \right) \quad \text{s.t.} \begin{cases} y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geqslant 1 - \xi_i \\ \xi_i \geqslant 0,\ C > 0 \end{cases}. \tag{5}$$