Information Sciences 370–371 (2016) 92–102

Contents lists available at ScienceDirect

Information Sciences

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

Tuning kernel parameters for SVM based on expected square distance ratio

Shen Yin*, Jiapeng Yin

Research Institute of Intelligent Control and Systems, Harbin Institute of Technology, Heilongjiang 150001, China

ARTICLE INFO

Article history:
Received 29 July 2015
Revised 14 July 2016
Accepted 20 July 2016
Available online 27 July 2016

Keywords:
Classification
Class separability
Kernel parameter
SVM

ABSTRACT

The performance of a support vector machine (SVM) depends highly on the selection of the kernel function type and relevant parameters. To choose the kernel parameters properly, methods analyzing the class separability have been widely adopted for their efficiency compared with other methods, such as the popular grid search algorithm. This paper proposes a novel index called the Expected Square Distance Ratio (ESDR), which can serve as a better class separability criterion than the existing ones. Experiments on real-world datasets show that, compared with common kernel parameter selection methods that utilize the between-class separation, the variations in ESDR with respect to the kernel parameter are much more in line with those of the classification accuracy, leading to better kernel parameters. Moreover, ESDR takes the exact data distribution into account and can thus be used to study the model selection problem of an SVM for certain forms of data distribution. As an example, we employ the ESDR to analyze the selection of RBF (Radial Basis Function) kernel parameters for Gaussian data classification.

1. Introduction

Since its development in the 1990s, a support vector machine (SVM) [32] has been widely used in the fields of pattern recognition [4,28,38,39,45] and regression estimation [10,17,33,41,44]. The performance of an SVM largely depends on the kernel function it adopts. Therefore, it is of great importance to choose the kernel type and set the corresponding parameters appropriately. However, to date, there have been no optimal methods that can lead to the correct kernel and its hyperparameters.

Despite this fact, there are still some effective approaches to selecting the proper model for an SVM, among which the grid search (GS) algorithm [16] is the most straightforward. When operating along with a cross-validation [15], GS can be quite effective and stable. However, GS suffers from a heavy computational burden because the SVM model has to be rebuilt for all combinations of parameters. To avoid searching the entire parameter space, some optimization algorithms, such as the genetic algorithm (GA) [7,26], particle swarm optimization (PSO) [31,43], simulated annealing (SA) technique [18,25], fruit fly optimization algorithm (FOA) [29], gravitational search algorithm (GSA) [24], social emotional optimization algorithm (SEOA) [46], firefly algorithm [5], and artificial chemical reaction optimization algorithm (ACROA) [1], can be applied to the SVM parameter tuning process. However, before the termination of such methods, the entire population has to be updated for many generations, and thus, they remain time-consuming processes. To solve this problem analytically, Chapelle et al.

* Corresponding author. Tel.: +8645186402350.
E-mail addresses: shen.yin@hit.edu.cn, shen.yin@uni-due.de (S. Yin), yjp.aaron@gmail.com (J. Yin).

http://dx.doi.org/10.1016/j.ins.2016.07.047
0020-0255/

[6] first proposed a generalization error estimation called a radius-margin bound, and used a gradient descent algorithm to arrive at good parameters. However, this method is only applicable to the L2-SVM and needs to solve an extra quadratic optimization problem, which makes it very time-consuming. Under Bayesian interpretations of an SVM, Gold et al. [12] chose the parameters by maximizing the available evidence. Gomes et al. [13] bonded the initial parameter pair obtained from meta-learning with search techniques such as a PSO, and employed a hybrid algorithm to select the hyperparameters. Under the assumption of a Gaussian distribution, Wang et al. [37] determined the optimal super-parameter of a Gaussian kernel that leads to sufficient support vectors before eliminating the outliers. Utilizing the distances from the samples to the enclosing surfaces, [42] derived an optimization problem to select a Gaussian kernel parameter for a one-class SVM.

To better take the data distribution into consideration and tune the SVM from the perspective of the geometry, the issue of class separability is introduced. Among all of the class separability criteria, scatter-matrix based measures [8] are extensively adopted for their simplicity. As a commonly used scatter-matrix based measure that can be interpreted in terms of a Fisher Discriminant Analysis [36], criterion J_4 has been successfully applied to solve the model selection problem of an SVM [34]. In addition, Sun et al. [30] proposed analyzing the distance between two classes (DBTC) to obtain the desired kernel parameters. Wu and Wang [40] presented the inter-cluster distance in the feature space, which is equivalent to DBTC. In this paper, we put forward a new index called the Expected Square Distance Ratio (ESDR), which can quantify the class separability well. Compared with other criterions such as DBTC, the ESDR has a clearer intuitive meaning and is more accordant with the accuracy of SVM classification when employed to select the kernel parameters for an SVM, illustrating that it reflects the geometric structures of the feature space corresponding to certain kernels. Moreover, explicitly taking the data distribution into account, the ESDR can serve as a powerful tool for the study of SVM model selection for certain forms of data distribution, such as a Gaussian distribution.

The penalty coefficient C also has a significant influence on the performance of an SVM. Methods such as a grid search technique and radius-margin bound incorporate C and kernel parameters into a unified framework [20]. However, the kernel parameters determine the geometry of the feature space, whereas C weighs the margin maximization and error minimization, and has no intuitive meaning in terms of geometry. Thus, the class separability is usually applied to determine the kernel parameters rather than C for an SVM. In our proposed method, the ESDR is employed first to select the optimal kernel parameters, and the penalty coefficient C is then determined through a cross-validation and grid search technique.

The rest of this paper is organized as follows. Section 2 reviews the theoretical background of an SVM. Section 3 analyzes the properties of the ESDR and compares them with other criteria, and applies the ESDR for the parameter tuning of an RBF kernel on Gaussian data. The results of several experiments conducted on some real-world datasets and Gaussian data are provided in Section 4. Finally, Section 5 provides some concluding remarks regarding this research.

2. Related work

SVMs have been extensively used in many fields with a good performance level [4,10,17,28,33,38,39,41,44,45]. Before the SVM emerged as an outstanding machine learning method, classical learning approaches such as neural networks [14] merely follow the empirical risk minimization (ERM) rule, the key point of which is minimizing the training error, leading to the problem of overfitting. Based on the Vapnik–Chervonenkis theory (VC-theory), an SVM follows the structural risk minimization (SRM) rule, which not only minimizes the training error but also restricts the complexity of the learning machine, thus improving the generalization abilities [32]. Owing to its well-established mathematical foundations, an SVM has been successfully applied to numerous learning tasks under various conditions.

The original SVM was proposed for binary classification. Considering the training set $(\boldsymbol{x}_i, y_i)$, $i = 1, 2, \ldots, n$, $\boldsymbol{x}_i \in \boldsymbol{R}^d$, where $\boldsymbol{x}_i$ is the training data vector of d-dimensions, $y_i \in \{-1, +1\}$ is the corresponding class label, and n is the size of the training set, the SVM constructs a separating hyperplane that results in zero training errors (assuming that the training set is linearly separable) and a maximal margin. The margin refers to the maximal width of the slab parallel to the hyperplane with no data points inside and where the optimal separating hyperplane is directly in the middle of the slab. For visualization, see Fig. 1 for the specific case of a two-dimensional space.

To obtain the optimal separating hyperplane $< \boldsymbol{\omega}, \boldsymbol{x} > + b = 0$, an optimization problem needs to be solved

$$\begin{aligned} &\min_{\boldsymbol{\omega}, b} \quad \frac{1}{2}\|\boldsymbol{\omega}\|^2 \\ &s.t. \quad y_i(< \boldsymbol{\omega}, \boldsymbol{x}_i > + b) \geq 1. \end{aligned} \tag{1}$$

For the sake of generalization, the separating hyperplane is occasionally allowed to not separate all of the training data. In this case, a soft margin is used, and the optimization problem (1) is transformed into

$$\begin{aligned} &\min_{\boldsymbol{w}, b, s} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_i s_i \\ &s.t. \quad y_i(< \boldsymbol{w}, \boldsymbol{x}_i > + b) \geq 1 - s_i, s_i > 0 \end{aligned} \tag{2}$$

where s_i is called a slack variable, and C is the penalty coefficient.

In a real-world task, however, data are rarely linearly separable. To solve the problem of nonlinearity, the original data should be projected through nonlinear mapping $\Phi(\boldsymbol{x})$ onto a new space, usually with higher dimensions, where the data points are linearly separable.