Applied Mathematics and Computation 218 (2012) 4973–4987

Contents lists available at SciVerse ScienceDirect

Applied Mathematics and Computation

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

Multi-class support vector machine optimized by inter-cluster distance and self-adaptive deferential evolution

Xiaoyuan Zhang, Jianzhong Zhou *, Changqin Wang, Chaoshun Li, Lixiang Song

College of Hydropower and Information Engineering, Huazhong University of Science and Technology, Wuhan 430074, PR China

ARTICLE INFO

Keywords:
Support vector machine
Parameter optimization
Inter-cluster distance
Differential evolution
Fault diagnosis
Rolling element bearings

ABSTRACT

Support vector machine (SVM) is a popular tool for machine learning task. It has been successfully applied in many fields, but the parameter optimization for SVM is an ongoing research issue. In this paper, to tune the parameters of SVM, one form of inter-cluster distance in the feature space is calculated for all the SVM classifiers of multi-class problems. Inter-cluster distance in the feature space shows the degree the classes are separated. A larger inter-cluster distance value implies a pair of more separated classes. For each classifier, the optimal kernel parameter which results in the largest inter-cluster distance is found. Then, a new continuous search interval of kernel parameter which covers the optimal kernel parameter of each class pair is determined. Self-adaptive differential evolution algorithm is used to search the optimal parameter combination in the continuous intervals of kernel parameter and penalty parameter. At last, the proposed method is applied to several real word datasets as well as fault diagnosis for rolling element bearings. The results show that it is both effective and computationally efficient for parameter optimization of multi-class SVM.

1. Introduction

Support vector machine (SVM) is based on the structural risk minimization (SRM) principle [1], which makes it less prone to over-fitting. By maximizing the margin between two opposite classes, SVM can find the optimal separating hyper-plane that minimizes the upper bound of the generalization error, which enables SVM to have strong capability of fitting and generalization. By introducing the kernel tricks, SVM has the ability of dealing with infinite or nonlinear features in a high dimensional feature space. With the above attractive features, SVM is regarded as state-of-the-art classifier. It is generally acknowledged that SVM has a good performance in solving nonlinear and high dimensional pattern recognition problems with good generalization ability. Although SVM has so many advantages and has been successfully applied in many fields, such as biomedicine [2,3], text categorization [4,5] fault diagnosis [6,7] and so on, in practice, its parameters, the kernel parameters (for instance, width parameter g of RBF kernel function) and penalty parameter C, must be selected judiciously so that the performance of SVM can be brought into full play. Changing the kernel parameters is equivalent to selecting the feature spaces, and tuning C is corresponding to weighting the slack variables, the error terms. Consequently, the performance of SVM depends on its parameters largely. However, there is no systematic methodology or priori knowledge for determining the parameters of SVM.

A wide range of studies have been carried out on this topic. A simple and straightforward way is grid search (GS) [8]. This procedure requires a grid search over the parameter space. It trains SVMs with all desired combinations of parameters and

* Corresponding author.
E-mail address: jz.zhou@mail.hust.edu.cn (J. Zhou).

doi:10.1016/j.amc.2011.10.063

screens them according to the training accuracy. It makes the training process time-consuming, and when the number of parameters exceeds two it will become intractable. Another approach is to build estimates or bounds for the true generation error, and to use numerical optimization methods [9,10] to minimize some analytical criterions which are proxies of these estimates or bounds. The numerical methods are generally more efficient than GS, for their fast convergence rate. However, they are sensitive to the initial point. If the initial point is not proper, the numerical methods might not properly solve the parameter optimization problem. Additionally, these methods require that the kernel functions and the bounds of generalization error have to be differentiable with respect to kernel and penalty constant parameters. Recently, some evolutionary algorithms (EAs), such as genetic algorithm (GA) [11,12], particle swarm optimization algorithm (PSO) [13,14], artificial immunization algorithm (AIA) [6,15] and ant colony optimization algorithm (ACO) [16] have been adopted to optimize the SVM parameters for their better global search abilities. These EAs provide an alternative for finding global optimal solutions in non-convex, highly nonlinear, noisy, time-dependent or flat solution spaces. Consequently, they can find the optimal parameters combination for SVM with a high probability. However, these methods select the best parameter combination from the population evolved generation by generation, which requires training many SVMs. Thus, they are still time-consuming, especially when the search ranges of the parameters are large.

SVM uses kernel functions to map the input data into a high dimensional feature space. Hence, the feature space is determined by the kernel function and its parameters. When the kernel function is selected, the feature space is only determined by the kernel parameters. Selecting kernel parameters is equivalent to selecting the feature space. Obviously, we need a feature space in which the classes are more separated. The inter-cluster distances in the feature spaces (ICDF) indicate the separation of two classes in the feature space. The larger the ICDF, the easier to classify the two classes in feature space. Wu and Wang [17,18] used ICDF to choose the kernel parameters. For binary classification problem, the optimal kernel parameters are found according to the largest ICDF. Then, the selected kernel parameters with different candidate penalty parameter C are used to train SVM models. The parameters combination which results in a highest cross-validation accuracy is selected as the best one. Since calculating ICDFs does not require the information of the trained SVMs, the time needed for the training process for different kernel parameters is saved. This method can get the parameter combinations which result in SVM models perform as good as the models chosen by traditional GS method in testing accuracy, while the training time is shortened largely. However, in this method, the candidate penalty parameters and kernel parameters for calculating ICDFs are given as discretization. It needs to locate the interval of feasible solution and a suitable sampling step, which is a tricky task since a suitable sampling step varies from kernel to kernel and the grid interval may not be easy to locate without prior knowledge of the problem [19]. Furthermore, in practice, most of the classification problems are multi-class, how to extend the binary SVM with parameter optimization by ICDF to multi-class problems is a challenging task.

In this paper, a hybrid method of ICDF index and self-adaptive differential evolution (ADE) algorithm (ICDF–ADE) is proposed to optimize the parameters of SVM. The proposed method capitalizes on the strengths of both the ICDF heuristic and ADE strategy. Firstly, for each class pair of a multi-class problem, the ICDFs are calculated, and the optimal kernel parameter for the SVM of this class pair is found according to the largest ICDF. Then, a small and effective search interval of kernel parameter covering the optimal kernel parameter of each class pair is determined. At last, ADE is used to search the optimal parameter combination of SVM in the continuous intervals of kernel parameter and penalty parameter. Since the search range of kernel parameter is much smaller than that of traditional methods, the training time of the proposed method is much shortened. Moreover, differential evolution (DE) has shown performance superior to that of PSO and other EAs in the widely used benchmark problems [20] and has fewer parameters to set. Thus, it outperforms other methods in optimizing the parameters of SVM. The proposed method is tested on several real word datasets as well as fault diagnosis for rolling element bearings. The experimental results show that it has good performance in optimizing the parameters of multi-class SVM, and thus is suitable for fault diagnosis of rolling element bearings.

The remaining of this paper is organized as follows. The brief introduction of SVM is presented in Section 2. Several forms of ICDF are given in Section 3. The basic ideas of ADE are introduced in Section 4. The proposed method as well as numerical experiments is described in detail in Section 5. In Section 6, the proposed method is applied in fault diagnosis for rolling element bearings and the experimental results are compared. Finally, a general conclusion is drawn in Section 7.

2. Support vector machines

2.1. Brief introduction of support vector machines

Support vector machines (SVM) were first suggested by Vapnik [1]. The principles of SVM stem from statistical learning theory. By using the information of limited samples, SVMs search for a compromise between the model complexity and learning ability to obtain good generalization ability. In this section, some basic conceptions for SVMs are introduced. For a more detailed discussion of SVM can be found in Cristianini and Shawe-Taylor [21].

Given a dataset $(\overrightarrow{x_i}, y_i), i = 1, \ldots, l, \overrightarrow{x_i} \in R^d, y_i \in \{1, -1\}$, where R^d is the d-dimensional input space, l is the number of training samples, $\overrightarrow{x_i}$ is the ith training sample and y_i is its corresponding bipolar label. A linear decision surface can be defined by the equation $f(\vec{x}) = \langle \vec{w}, \vec{x} \rangle + b = 0$, where $\vec{w}$ is a weight vector orthogonal to the decision surface, b is an offset term, $\langle \cdot, \cdot \rangle$ indicates the inner product operation. The original formulation of SVM algorithm seeks a linear decision surface that separates