Expert Systems with Applications 38 (2011) 3227–3237

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Model selection for least squares support vector regressions based on small-world strategy

Wentao Mao *, Guirong Yan, Longlei Dong, Dike Hu

The Key Laboratory of Strength and Vibration of Ministry of Education, Xi'an Jiaotong University, Xi'an, China

ARTICLE INFO

Keywords:
Model selection
Least squares support vector machines
Small-world
Tabu search

ABSTRACT

Model selection plays a key role in the application of support vector machine (SVM). In this paper, a method of model selection based on the small-world strategy is proposed for least squares support vector regression (LS-SVR). In this method, the model selection is treated as a single-objective global optimization problem in which generalization performance measure performs as fitness function. To get better optimization performance, the main idea of depending more heavily on dense local connections in small-world phenomenon is considered, and a new small-world optimization algorithm based on tabu search, called the tabu-based small-world optimization (TSWO), is proposed by employing tabu search to construct local search operator. Therefore, the hyper-parameters with best generalization performance can be chosen as the global optimum based on the powerful search ability of TSWO. Experiments on six complex multimodal functions are conducted, demonstrating that TSWO performs better in avoiding premature of the population in comparison with the genetic algorithm (GA) and particle swarm optimization (PSO). Moreover, the effectiveness of leave-one-out bound of LS-SVM on regression problems is tested on noisy sinc function and benchmark data sets, and the numerical results show that the model selection using TSWO can almost obtain smaller generalization errors than using GA and PSO with three generalization performance measures adopted.

1. Introduction

As an important branch of machine learning, least squares support vector machine (LS-SVM), introduced by Suykens and Vandewalle (1999), has been a promising tool in the fields of character recognition, signal processing and stock action prediction, etc. (Baylar, Hanbay, & Batan, 2009; Hanbay, 2009; Kumar & Gopal, 2009; Shin, Lee, & Kim, 2005) in the last decade. Model selection is a key issue in LS-SVM applications. There are two hyper-parameters, i.e. regularization parameter and kernel parameter, in LS-SVM model. In general, the cross validation (CV) errors are widely used as the generalization performance measure to control the selection of hyper-parameters of regression model (Duan, Keerthi, & Poo, 2003). Two typical approaches of this method are 5-fold CV and leave-one-out (LOO) methods. Though these methods have advantages such as simplicity, reliability, etc., they are computationally expensive. Chang and Lin (2005) derived theoretically various LOO bounds for $\varepsilon - SVR$ which is introduced by Vapnik (1995) and discussed continuity and differentiability of LOO bounds, e.g. Radius-Margin Bound and Span Bound, using $L2 - SVR$. These bounds are easy to compute and these two properties supply a theoretical foundation to select optimal parameters using the gradient-based optimization methods. But the model selection for LS-SVM, especially for LS-SVM regression (LS-SVR) problem, has not been studied widely. One of the reasons is replacing the quadratic programming with equality constraints to solving a set of linear equations in LS-SVM. As a result, the existing analytical bounds of classical support vector machines (SVMs) cannot be adopted directly for LS-SVMs. Guo, Yang, Wu, Wang, and Liang (2008) utilized 5-fold CV to carry out model selection of LS-SVM classifications, but high computational cost is unavoidable. Cawley and Talbot (2007) used matrix manipulation to derive a closed form of LOO bound and introduced skillfully the Bayesian method to solve the potential over-fitting problem produced by LOO method. This bound is just a byproduct of single training procedure. Therefore, LOO bound is a simple and mathematically tractable criterion with very low computational expense. However, the effectiveness of this method is tested only for the classification problems and no recommendations are made in the literature on how the LOO bound can be applied to LS-SVM regression.

Because the task is to pursue the best generalization performance, model selection can be essentially considered as a single-objective global optimization problem where the generalization performance measure performs as fitness function. In practice, the widely used strategy is grid search over the whole parameter

* Corresponding author. Tel.: +86 29 82660978; fax: +86 29 82664849.
E-mail address: maowt.mail@gmail.com (W. Mao).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.08.109

space. It is unavoidably time consuming since this method requires retraining the model many times under different parameter settings. Another strategy is to minimize the theoretical error bounds using iterative gradient descent techniques. This strategy needs the bound functions are differentiable. Although less computational burden is needed, the choice of initial values will heavily affect the final results. The intelligence algorithms represented by genetic algorithm (GA) (Michalewicz, 1996) and particle swarm optimization (PSO) (Kennedy & Eberhard, 1995) can find the global optima effectively, therefore they have been successfully applied to model selection (Avci, 2009; Guo et al., 2008). However, these heuristic algorithms usually fall into premature of the population when solving complex optimization problems and thus obtain results with low precision. Therefore, the performance of model selection could be improved after the deficiency discussed above has been corrected.

As a new optimization method based on the model of complex small-world network, the small-world optimization algorithm (SWA) (Du, Shao, & Feldman, 2006) can find the global optima more effectively. However, the traditional small-world optimization algorithms, proposed in Du, Wu, and Zhuang (2006) and Wang, Yeung, and Lochovsky (2008), adopt binary encoding strategy. In Du and Shao et al. (2006), the decimal-coding SWA needs to be converted from decimal to binary every single time the cost function is calculated. The decimal encoding strategy can increase precision as there will be no loss in resolution during coding, and the problem space can be searched more thoroughly and rigorously. Michalewicz (1996) proved that the performance of decimal encoding strategy is better than binary encoding no matter from the perspective of convergence speed or result's accuracy. Another advantage of decimal encoding is the gradient information could be utilized directly. In this paper, a new decimal-coding small-world optimization algorithm based on tabu search, called the tabu-based small-world optimization (TSWO), is proposed by employing tabu search to construct local search operator. Compared with GA and PSO, TSWO has better performance of global optimization and further provide the possibility of adopting gradient information which has been computed in Chang and Lin (2005) for $\varepsilon - SVR$ and (Cawley, 2006) for LS-SVR, respectively. Therefore, TSWO is better suited to model selection.

This paper is organized as follows. In Section 2, we provide a brief review to traditional SWA and tabu search algorithm. In Section 3, the idea and design of TSWO are elaborated in detail. Adopting LOO bound proposed in Cawley (2006) as the fitness function of TSWO, a new model selection method based on TSWO for LS-SVR is presented in Section 4. Experimental results on multimodal function and benchmark regression data sets are then presented in Section 5, followed by a conclusion of the paper in last section.

2. Small-world algorithm and tabu search

2.1. Small-world optimization algorithm

Rooted from the research on "tracking the shortest paths in American social networks" in 1960's (Wang, Yang, & Jiao, 2006), the small-world phenomenon, also called "six degrees of separation", indicates that on average any two persons in the world could be linked by six acquaintances (Du & Shao et al., 2006). This can be regarded as an efficient mechanism of transferring information which depends on dense local connections and a few long-range connections to find the shortest paths to reach the destination. Furthermore, Kleinberg (2001) pointed out that the optimal paths could be found only by using local information. Inspired by these conclusions, Du and Wu et al. (2006) treated the solution space of optimization problem as a small-world network and looked the optimization procedure as finding the shortest paths from candidate solution to optimal solution. Characterized by the small-world effect, a small-world optimization algorithm was developed in Du and Wu et al. (2006).

Du and Shao et al. (2006) gives a detailed description of SWA. Take a two-dimensional optimization problem for example. Suppose the problem as:

$$\begin{array}{ll} \text{minimize} & f(x_1, x_2) \\ \text{s.t.} & d_i \leqslant x_i \leqslant u_i \end{array}$$

where $\boldsymbol{d} = \{d_1, d_2\}$, $\boldsymbol{u} = \{u_1, u_2\}$ are the lower and upper bounds of corresponding variable. Fig. 1 shows the feasible solution space of $x1 - x2$. With pre-defined precision, the values of candidates are represented by the circles. As shown in Fig. 1, T is the optimal solution and B is one initial candidate. The solid line in Fig. 1 illustrates the optimization procedure, in which global optimal solution can be found mainly by dense local search and a few long-range connections.

The realization of SWA will be briefly described as follows. To improve the efficiency, a population of candidates instead of a single candidate is defined as the starting point of SWA. The binary encoding strategy is adopted. The key of SWA includes a local search operator Ψ and a random long-range search operator Γ. The function of Ψ transfers information from node $s_i(k)$ to node $s_i(k+1)$ which is nearest to the global optimal solution with smaller value of $\|s_i(k+1) - s_i(k)\|$. A number of bits are randomly chosen from the whole binary encoding of $s_i(k)$ to construct local subnet and the bit-oriented adding, minus, multiply on these bits represent transmission of local information. According to the predefined probability, Γ stochastically chooses a node $s_i'(k)$ in $non - l$ neighborhood set with bigger value of $\|s_i'(k) - s_i(k)\|$. In realization, the Inverse Operation in evolutionary computation is used to construct operator Γ. The search then progresses by iteratively moving from the initial population to the optimal solution, which is also illustrated in Fig. 1. Compared with GA and evolutionary algorithm, SWA does not include crossover operation, and the global search operation is much less important than local selection operation. Moreover, Du and Shao et al. (2006) gave very detailed comparisons of SWA and GA.

2.2. Tabu search

Belonging to the iterative neighborhood search methods, Tabu search (TS), introduced in Glover (1989), tries to solve hard optimi-

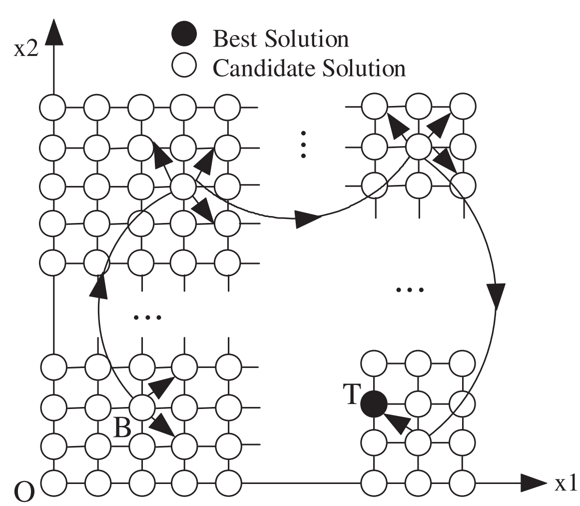

Fig. 1. Illustration of optimization procedure characterized by small-world phenomenon in a two-dimensional solution space.