Knowledge-Based Systems 96 (2016) 61–75

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Evolving support vector machines using fruit fly optimization for medical data classification

Liming Shen[a], Huiling Chen[a,c,*], Zhe Yu[a], Wenchang Kang[a], Bingyu Zhang[a], Huaizhong Li[a], Bo Yang[b,c], Dayou Liu[b,c]

[a] College of Physics and Electronic Information Engineering, Wenzhou University, 325035 Wenzhou, China
[b] College of Computer Science and Technology, Jilin University, Changchun 130012, China
[c] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China

ARTICLE INFO

Article history:
Received 7 March 2015
Revised 31 December 2015
Accepted 2 January 2016
Available online 11 January 2016

Keywords:
Support vector machine
Parameter optimization
Fruit fly optimization
Medical diagnosis

ABSTRACT

In this paper, a new support vector machines (SVM) parameter tuning scheme that uses the fruit fly optimization algorithm (FOA) is proposed. Termed as FOA-SVM, the scheme is successfully applied to medical diagnosis. In the proposed FOA-SVM, the FOA technique effectively and efficiently addresses the parameter set in SVM. Additionally, the effectiveness and efficiency of FOA-SVM is rigorously evaluated against four well-known medical datasets, including the Wisconsin breast cancer dataset, the Pima Indians diabetes dataset, the Parkinson dataset, and the thyroid disease dataset, in terms of classification accuracy, sensitivity, specificity, AUC (the area under the receiver operating characteristic (ROC) curve) criterion, and processing time. Four competitive counterparts are employed for comparison purposes, including the particle swarm optimization algorithm-based SVM (PSO-SVM), genetic algorithm-based SVM (GA-SVM), bacterial forging optimization-based SVM (BFO-SVM), and grid search technique-based SVM (Grid-SVM). The empirical results demonstrate that the proposed FOA-SVM method can obtain much more appropriate model parameters as well as significantly reduce the computational time, which generates a high classification accuracy. Promisingly, the proposed method can be regarded as a useful clinical tool for medical decision making.

1. Introduction

As a primary machine learning paradigm, support vector machines (SVM) [1,2] are rooted in the Vapnik–Chervonenkis theory and the structural risk minimization principle. The SVM attempts to determine a tradeoff between minimizing the training set error and maximizing the margin in order to achieve the best generalization ability and remain resistant to over fitting. Additionally, one major advantage of the SVM is the use of convex quadratic programing, which provides only global minima; thus, it avoids being trapped in local minima. Due to its advantageous nature, SVM has been applied to a wide range of classification tasks [3–6]. In particular, SVM has been shown to perform very well on many medical diagnosis tasks [5–11]. However, there is still a need for improving the SVM classifier's performance.

It has been demonstrated that the SVM classification accuracy can be substantially improved by establishing the proper model parameter settings [12]. Thus, key parameters, such as the penalty parameter and the kernel bandwidth of the kernel function, should be properly determined prior to its application to practical problems. The first parameter, the penalty parameter C, determines the trade-off between the fitting error minimization and the model complexity. The second parameter, the kernel bandwidth γ, defines the non-linear mapping from the input space to some high-dimensional feature space. Traditionally, these parameters were handled by the grid-search method [13] and the gradient descent method [14–16]. However, these methods are vulnerable to local optimum. Recently, biologically-inspired metaheuristics (such as the genetic algorithm [17], particle swarm optimization (PSO) [18], and bacterial foraging optimization (BFO) [19]) have been shown to be more likely to determine the global optimum solution than the traditional aforementioned methods.

Not only are the traditional optimization algorithms, such as GA, BFO, and PSO, complex to implement, but they are also difficult to understand. In addition, determining the global optimal solution is time-consuming. As a new member of the swarm-intelligence algorithms, the fruit fly optimization algorithm (FOA) [20] is inspired by the foraging behavior of real fruit flies. The FOA has certain outstanding merits, such as a simple computational process,

* Corresponding author at: College of Physics and Electronic Information Engineering, Wenzhou University, 325035 Wenzhou, China. Tel.: +86057786689125.
E-mail address: chenhuiling.jlu@gmail.com (H. Chen).

http://dx.doi.org/10.1016/j.knosys.2016.01.002
0950-7051/

simple implementation, and easy understanding with only a few parameters for tuning. Due to its good properties, FOA has become a useful tool for many real-world problems. In 2012, Li et al. [21] proposed the FOA to optimize the parameters in LSSVM (LSSVM-FOA) and applied it to forecast the annual electric load. The computational result showed that the proposed method outperformed the other alternative methods. In 2013, Chen et al. [22] proposed using the FOA to tune the parameters in the gray model neural network and applied the resultant model, FOAGMNN, to construct service satisfaction detection. Based on the experimental analysis results, the FOAGMNN model achieved the fastest error convergence and the best classification capability. In 2013, Li et al. [23] developed a hybrid annual power load forecasting model that combined FOA and generalized regression neural network (GRNN). In the proposed method, FOA was used to determine the appropriate spread parameter in GRNN, and the effectiveness of this proposed hybrid model was demonstrated in two experiment simulations. Both experiments revealed that the proposed hybrid model outperformed the GRNN model with the default parameter and others. In 2013, Shan et al. [24] presented an improved FOA (LGMS-FOA) that enhanced the FOA's performance. LGMS-FOA's superiority was demonstrated by simulation results and by comparing LGMS-FOA to FOA and other metaheuristics. Both the searching efficiency and the searching quality were greatly improved. In 2013, Wang et al. [25] proposed a novel binary fruit fly optimization algorithm (bFOA) to solve the multidimensional knapsack problem (MKP). In the bFOA, several techniques were introduced, including the binary string, local and global vision-based search process, a group generating probability vector, a global vision mechanism, two repair operators, and the Taguchi method. A comparative study demonstrated that the proposed bFOA effectively solved the MKP, especially for large-scale problems. Recently, Pan [26] proposed a modified FOA (MFOA), in which an escape parameter is added to the distance function. The findings showed that the forecasting model that combines MFOA and GRNN has the best ability for forecasting the closing price of both oil and gold. Pan and colleagues [27] presented an improved fruit fly optimization (IFFO) algorithm that introduced a new control parameter to adaptively tune the search scope around its swarm location for solving continuous function optimization problems. Yuan et al. [28] proposed a variation on the original FOA technique, named the multi-swarm fruit fly optimization algorithm (MFOA), by employing multi-swarm behavior to significantly improve the performance. By applying the proposed MFOA approach to several benchmark functions and the parameter identification of a synchronous generator, it was demonstrated that the proposed approach is superior to the original FOA technique. Li et al. [29] presented an improved fruit fly optimization algorithm (FOA) via a well-designed smell search procedure to solve the steelmaking casting problem, and the simulation results indicate that the proposed FOA is more effective than the four presented algorithms. Zheng et al. [30] proposed a novel fruit fly optimization algorithm (nFOA) based on multiple fruit fly groups to solve the semiconductor final testing scheduling problem, and the experimental results demonstrated that the nFOA effectively and efficiently solved the SFTSP. More recently, Wang et al. [31] presented an effective and improved FOA (IFOA) for optimizing numerical functions and solving joint replenishment problems. The improvements include a new method for maintaining the population diversity in order to enhance the exploration ability, a new parameter to avoid local optimization, and a random perturbation to the updated initial location to jump out of the local optimum. Furthermore, a comparative study reveals that the proposed IFOA can obtain better solutions than the current best algorithm. Wang et al. [32] proposed an adaptive mutation fruit fly optimization algorithm (AM-FOA) to optimize the parameters in the least squares support vector machine (LSSVM), and they successfully applied the resultant model, AM-FOA-LSSVM, to the practical melt index prediction problem. Xiao et al. [33] presented an improved FOA based on the cell communication mechanism (CFOA) that takes into consideration the global worst, mean, and best solutions into the search strategy to improve the exploitation. The results from a set of numerical benchmark functions show that the CFOA outperforms the FOA and the PSO in most experiments. Furthermore, the CFOA is applied to optimize the controller of peroxidation furnaces in carbon fibers production, and simulation results have validated the CFOA's effectiveness.

From the above FOA-related works, FOA has been proven to be effective for parameter optimization for machine learning algorithms, including the GRNN and LSSVM. Additionally, it should be noted that the related LSSVM parameter optimization based on FOA is only suitable for the regression problem; it cannot be used for classification problems. Therefore, this study will be the first to explore the FOA technique's ability to address SVM's model selection problem for classification. Furthermore, the resultant model, FOA-SVM, successfully and effectively detected the medical data classification problem. In the proposed FOA-SVM method, the cross validation classification accuracy is considered for designing the objective function to explore SVM's maximum generalization ability. The proposed method's effectiveness and efficiency was examined in terms of the classification accuracy, sensitivity, specificity, AUC, and CPU time on the medical datasets taken from the UCI machine learning repository. As the experimental results will show, our proposed method can obtain more appropriate model parameters and obtain a high predictive accuracy with much less processing time as compared to the grid search-based and other metaheuristic-based methods.

The remaining of this paper is organized as follows. Section 2 gives some brief background knowledge of SVM and FOA. The detailed implementation of the FOA-SVM methodology will be explained in Section 3. Section 4 describes the experimental design in detail. The experimental results and discussions of the proposed approach are presented in Section 5. Finally, conclusions are summarized in Section 6.

2. Background materials

2.1. Support vector machines (SVM)

This section gives a brief description of SVM. For more details, one can refer [2,34], which provides a complete description of the SVM theory.

Let us consider a binary classification task: $\{xi, yi\}$, $i = 1, \ldots, l$, $yi \in \{-1, 1\}$, $xi \in R^d$, where xi are data points, and yi are the corresponding labels. They are separated with a hyper plane given by $w^T x + b = 0$, where w is a d-dimensional coefficient vector that is normal to the hyper plane, and b is the offset from the origin. The linear SVM obtains an optimal separating margin by solving the following optimization task:

$$\begin{aligned} &\mathrm{Ming}(\mathrm{w},\xi) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ &s.t.\ ,\ y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \xi_i \geq 0 \end{aligned} \tag{1}$$

By introducing Lagrangian multipliers $\alpha_i (i = 1, 2, \ldots, n)$, the primal problem can be reduced to a Lagarangian dual problem:

$$\begin{aligned} &\max_{\alpha} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j \\ &s.t.\ ,\ \ \alpha_i \geq 0, i = 1, \ldots, n,\ \sum_{i=1}^{n}\alpha_i y_i = 0 \end{aligned} \tag{2}$$