Chemometrics and Intelligent Laboratory Systems 147 (2015) 147–156

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

A new data classification method based on chaotic particle swarm optimization and least square-support vector machine

Fang Liu [a,b,c,d,*], Zhiguang Zhou [a]

[a] Institute of information, Zhejiang University of Finance & Economics, Hangzhou 310018, China
[b] The State Key Laboratory of Mechanical Transmissions, Chongqing University, Chongqing 400044, China
[c] Guangxi Key Laboratory of Hybrid Computation and IC Design Analysis, Guangxi University for Nationalities, Nanning 530006, China
[d] The Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou 215006, China

ARTICLE INFO

Article history:
Received 3 June 2015
Received in revised form 28 July 2015
Accepted 17 August 2015
Available online 23 August 2015

Keywords:
Chaotic optimization algorithm
Particle swarm optimization algorithm
Least square-support vector machine (LS-SVM)
Data classification
Generalization ability
Classification accuracy

ABSTRACT

In order to improve the classification accuracy in chemometrics data, chaotic optimization algorithm (COA) and particle swarm optimization (PSO) algorithm are introduced into least square-support vector machine (LS-SVM) model in order to propose an optimized LS-SVM model and a novel data classification (CPL-SVM) method in this paper. In the proposed CPL-SVM method, the COA with the randomness and ergodicity is used to chaotically process the initial position and local extreme position of particle in the PSO algorithm in order to obtain a chaotic particle swarm optimization (CPSO) algorithm, and the CPSO is used to select and optimize the important parameters of LS-SVM, then the optimized parameters are used to obtain a better CPL-SVM classification method. The choice randomness of parameters is avoided and the selection workload of parameters is reduced. And this method can not only overcome the time-consuming and blindness of cross validation method, but also reflect small sample learning ability. In order to verify the effectiveness of CPL-SVM method, binary classification data, IRIS flower data and three relevant data sets with pharmacodynamic properties of drug are selected in this paper. The experiment results show that the proposed CPL-SVM method takes on the better learning performance, strong generalization ability, best sensitivity, Matthews correlation coefficient and classification accuracy. And it can effectively avoid the isolated effects of sample in the learning process.

1. Introduction

Chemometrics has been defined as the chemical discipline that uses mathematical and statistical methods to design and select optimal procedures and experiments, and provide the maximum chemical information based on analyzing chemical data. The most prominent part of chemometrics is data classification by using some intelligent methods for all obtained data. The chemometrics has mainly involved the information extraction from these obtained data. The available data and desired information often exist the hidden relationship, the analysis goal of chemometrics is to find out some relationships and classify these data by using new intelligent methods. These intelligent methods include neural networks (NN), genetic algorithm (GA), simulated annealing (SA), particle swarm optimization (PSO) algorithm, statistical analysis, support vector machine (SVM) and so on. However, the data of chemometrics have mostly multi-factors, high noise, nonlinear and irregular, so these complex data are classified in order to discovery the interdependent relationships and extract data model among these features. The SVM model [1] is a machine learning method based on the minimum structural risk principle for data classification by Vapnik. This model is a convex optimization problem to find out the global optimization solution. It can effectively solve these complex problems with the small sample, nonlinear, local minimum, avoid the slow convergence speed, and easily fall into the local minimum value. On the basis of SVM model, Suykens et al. [2,3] proposed a least square-support vector machine (LS-SVM) method. The LS-SVM model is to transform the SVM from the quadratic programming problem into the linear equations in order to reduce the computational complexity and improve the calculation speed in processing large sample. The LS-SVM has been widely applied in the pattern recognition, data mining, image analysis, network security and so on. However, the optimization parameters of LS-SVM model have an important influence on its optimization performance and learning precision. So how to optimize parameters of LS-SVM model is an important research problem in machine learning.

Chaotic optimization algorithm (COA) with the randomness and ergodicity is introduced into the PSO algorithm in order to make up for the low convergence speed, the late time oscillation and easy falling local minimum value in this paper. And a chaotic particle swarm optimization (CPSO) algorithm based on combining the COA and PSO is proposed. In LS-SVM model, the regularization parameter γ and radial basis kernel width parameter σ are very important for the optimization performance. It is an open problem in the field of LS-SVM how to find

* Corresponding author. Tel.:+86 571 8755 7136.
E-mail address: puigpuig2010@gmail.com (F. Liu).

http://dx.doi.org/10.1016/j.chemolab.2015.08.015
0169-7439/

the optimal values of regularization parameter γ and radial basis kernel width parameter σ. The common selection method is the cross validation method, but this method will not only consume a lot of computing time, but also take on certain blindness. So the proposed CPSO algorithm with global optimization ability is used to optimize the parameters of LS-SVM model. This can not only overcome the time-consuming and blindness of the cross validation method, but also reflect small sample learning ability, so as to improve the learning performance, generalization ability and robustness. Finally, a new data classification method based on the CPSO algorithm and LS-SVM model (CPL-SVM) is proposed in this paper. The binary classification data, IRIS flower data and three relevant data sets with pharmacodynamic properties of drug are selected to verify the effectiveness of the proposed CPL-SVM method.

The rest of this paper is organized as follows. Section 2 briefly introduces the related works about SVM, LS-SVM and their improved methods in the classification. Section 3 briefly introduces the related basic methods, including the COA, PSO algorithm, LS-SVM model and diversity-guided mutation strategy. Section 4 presents a chaotic particle swarm optimization algorithm, named the CPSO algorithm. Section 5 presents a novel data classification (CPL-SVM) method. In this section, the thoughts, model and the steps of the CPL-SVM method are introduced in detail. Section 6 applies and analyzes the CPL-SVM method in solving data classification problem. Finally, the conclusions are discussed in Section 7.

2. Related works

In recent years, in allusion to the optimization parameters of the SVM model or LS-SVM model, many researchers have deeply studied and explored from the different views in optimizing these parameters. They have proposed some optimization methods of parameters of the SVM model or LS-SVM model, such as empirical selection method, gradient descent method, cross validation method, GA, PSO algorithm and so on [4–8]. Temkoa et al. [9] proposed a fuzzy integral based combining different information sources to classify a small set of highly confusable human non-speech sounds. Devos et al. [10] proposed a methodological approach to guide the optimization parameters of SVM based on a grid search for minimizing the classification error rate. Tao et al. [11] proposed a new fast pruning algorithm for chemical pattern classification. Ghorbanzad'e and Fatemi [12] proposed a classification method of central nervous system agents by using LS-SVM based on their structural descriptors. Li et al. [13] proposed a novel automatic speaker age and gender identification approach based on combining seven different methods in order to improve the baseline performance. Huang et al. [14] proposed an informative novel tree kernel SVM classifier to model the relationship between bioactivity and molecular descriptors. Dong and Luo [15] proposed a new method to achieve bearing degradation classification based on principal component analysis (PCA) and optimized LS-SVM method. Lou'i et al. [16] proposed two new multisensor data fusion algorithms to reduce the rate of false detection and obtain reliable decisions on the presence of target objects. Zhang [17] proposed an improved data classification method based on SVM applying rational sample data selection and GA-controlled training parameters optimization. Yao and Yi [18] proposed a new License Plate (LP) detection technique based on multistage information fusion. Sung and Chung [19] proposed a distributed energy monitoring network system based on data fusion via improved PSO algorithm. He et al. [20] proposed a new method for classifying electronic nose data in rats wound infection detection based on SVM and wavelet analysis. Subhajit et al. [21] proposed a PSO method along with adaptive K-nearest neighborhood based gene selection technique to distinguish a small subset of useful genes.

For the optimization parameters of SVM model and LS-SVM model, although these scholars have done the in-depth study and discussion by using the various optimization methods from the different angle degree in order to obtain some good results, each proposed method has its own defect in optimizing parameters of SVM model and LS-SVM model, such as the low classification accuracy, weak generalization ability, slow convergence speed, and so on. So the CPSO algorithm based on COA and PSO is proposed to select and optimize the parameters of LS-SVM model in order to improve the classification accuracy, learning performance and generalization ability.

3. Basic methods

3.1. Chaotic optimization algorithm (COA)

Chaos often exists in the nonlinear system. It is a kind of characteristic that has a bounded unstable dynamic behavior and exhibits sensitive dependence on the initial conditions. Chaotic optimization algorithm (COA) [22] is a population-based stochastic optimization algorithm by using the chaotic mapping. The basic procedure of the COA is divided into two steps. First, the COA searches all the points in turn within the changing range of variables, and selects the better point as the current optimum point by using chaotic ergodicity, regularity, initial sensitivity and topological transitivity. Then the current optimum point is regarded as the center, a tiny chaotic disturbance is imposed and a careful search is performed in order to find out the global optimum point with the higher probability. Due to the chaotic non-repetition, the COA can carry out the overall search with the higher speed. So the COA takes on the characteristics of the easy implementation, short execution time and robust mechanism.

Currently, there have been several kinds of the COA based on chaotic characteristics, such as adaptive mutative scale COA [23], a mutative scale COA [24], chaotic harmony search algorithm [25], multi-objective chaotic ant swarm optimization [26] and so on. Because the adaptive mutative scale COA has the refined search space, better search speed and higher search accuracy [23], it is used to optimize the particle swarm optimization (PSO) algorithm in this paper. Generally, the main problem of the COA is to obtain chaotic variables. So the Logistic chaotic model is used to generate the chaotic variable. The mapping equation of the Logistic model is described:

$$Z_{n+1} = L(\mu, X_n) = \mu Z_n(1-X_n)\mu \in [0, 4], n = 0, 1, 2, 3, \cdots \quad (1)$$

where control variable ($\mu \in [0, 4]$) is the parameter of the Logistic. It has shown, when $Z_n \in [0, 1]$, the Logistic mapping is in the chaotic state. That is, the generated sequences under Logistic mapping function (the initial condition Z_0) are not periodic and converge. But the generated sequences must converge to one specific value outside the given range.

3.2. Particle swarm optimization (PSO)

The PSO algorithm [27] is a search algorithm based on simulating the social behavior of birds within a flock. In the PSO algorithm, individuals, referred to as particles, are "flown" through hyper dimensional search space. The positions of particles within the search space are changed based on the social-psychological tendency of individuals in order to delete the success of other individuals. The changing of particle within the population is influenced by the experience, or knowledge. The consequence of modeling for the social behavior is that the search is processed in order to return toward previously successful regions in the search space. Namely, the velocity (v) and position (x) of each particle will be changed according to the following expressions:

$$v_{ij}(t+1) = wv_{ij}(t) + c_1 r_1 \left(pB_{ij}(t) - x_{ij}(t)\right) + c_2 r_2 \left(gB_{ij}(t) - x_{ij}(t)\right) \quad (2)$$

$$x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1) \quad (3)$$

where $v_{ij}(t + 1)$ is the velocity of particle i^{th} at iteration j^{th}, $x_{ij}(t + 1)$ is the position of particle i^{th} at iteration j^{th}. w is inertia weight to be employed to control the impact of the velocity of previous history.