Journal of Universal Computer Science, vol. 23, no. 7 (2017), 603-618
submitted: 31/3/17, accepted: 30/6/17, appeared: 28/7/17 © J.UCS

Selecting Parameters of an Improved Doubly Regularized Support Vector Machine based on Chaotic Particle Swarm Optimization Algorithm

Chuandong Qin
(North Minzu University, Yinchuan, China
qcd369@163.com)

Zhenxia Xue
(Northern Michigan University, Marquette, USA
xuezhenxia@163.com)

Quanxi Feng
(Oklahoma State University, Stillwater, USA
fqx9904@163.com)

Xiaoyang Huang
(Xiamen University, Ximen, China
corresponding author: xyhuang@xmu.edu.cn)

Abstract: Taking full advantages of the L1-norm support vector machine and the L2-norm support vector machine, a new improved double regularization support vector machine is proposed to analyze the datasets with small samples, high dimensions and high correlations in the parts of the variables. A kind of smooth function is used to approximately overcome the disdifferentiability of the L1-norm and the steepest descent method is used to solve the model. But the parameters of this model bring some trouble about the accuracy of the experiments. By use of the characteristics of chaotic systems, we propose a chaotic particle swarm optimization to select the parameters in the model. Experiments show the improvement gains the better effects.

Key Words: L1 norm support vector machine, L2 norm support vector machine, chaotic particle swarm optimization, double regularization support vector machine
Category: H.2, I.2.11,I.5.2

1 Introduction

Based on the principle of structural risk minimization and the nature of Statistical Learning Theory (SLT), Support Vector Machine (SVM) was first proposed to deal with finite training datasets by Vapnik [Vapnik, 2013]. In order to prevent the over-fitting phenomenon, a Structural Risk Minimization (SRM) was used to process the finite training data sets. At the same time, a regularization or a penalty term was added to SRM, which based on the empirical risk function. It describes a general model of capacity control and provides a trade-off between hypothesis space complexity (the VC dimension of approximating functions) and the quality of fitting the training data (empirical error). Generally speaking, regularization about model complexity is a monotonically increasing function. That means the more complexity of the model, the greater the regularizer value. Regularization term can be the norm of the model parameter [Deng and Tian, 2004]. The model has the following form:

$$\min_{f \in F} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f) \tag{1}$$

Here the first term is the empirical risk which usually is replaced by some loss functions and the second term is the regularization which may be some vector parameter. $\lambda \geq 0$ is the parameter which controls the trade-off between the loss function $L(y_i, f(x_i))$ and the penalty $J(f)$. According to the maximum distance method for two classification hyperplanes [Shawe-Taylor and Cristianini, 2004], the regularization can select the L2-norm of the vector parameter and the model can be showed as the following:

$$\min_{f \in F} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) + \frac{\lambda}{2} \|w\|_2^2 \tag{2}$$

Here $\|w\|_2^2$ is the Euclidean Distance which also can be used in the ridge regression and neural network. The L2-norm penalty can help groups of correlated variables get selected together and it tends to make highly correlated input variables which have similar fitted coefficients [Li et al., 2006]. We often call them the grouping effect. With the help of this penalty, the number of selected input variables are no longer bounded by samples number n. However, $\|w\|_2^2$ can not produce the sparse coefficients and automatically select variables. On this occasion, some researchers proposed to replace $\|w\|_2^2$ with $\|w\|_1$. The L1-norm penalized optimization problem can be showed as the following optimization problem:

$$\min_{f \in F} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda \|w\|_1 \tag{3}$$

Contrary to the L2-norm, the numbers of variables which is selected by the L1-norm is upper bounded by the sample size n. For highly relevant variables, the L1-norm tends to select only one or few of them. L1-norm penalty has a unique property that is selecting variable automatically. Just as the two norms for their advantages and disadvantages, each coin has the two sides. Some researchers made full use of the advantages of the two norms and proposed a double regularization support vector machine [Li et al., 2016, Li et al., 2008]. The model

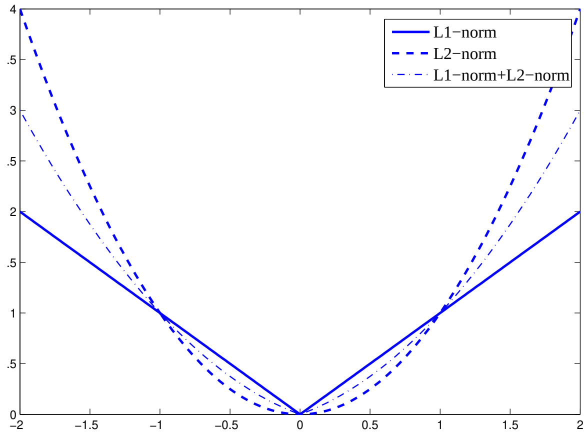

Figure 1: 2-dimensional contour plots of L1-norm, L2-norm and L1-norm+L2-norm, which have the different smoothness at the zero point.

can be expressed as the following form:

$$\min_{f \in F} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2 \tag{4}$$

The description of them can be found in Figure 1 in the two-dimensional space. The hybrid norm support vector machine for microarray classification is first proposed by Juntao [Li and Jia, 2010] and it has two major benefits:

◇ It can select variables automatically.

◇ It has the grouping effect, where highly correlated variables tend to be selected or removed together.

This model is very effective for classification, but the optimal solution of the problem (4) is very complex [Wu and Zhang, 2011]. In the following paper, we will introduce a polynomial smooth function and positive function to change the character approximately and give the steepest descent algorithm to gain the solutions. We choose the chaotic particle swarm optimization to select the parameters of the model. With regards to the above design principles, the contributions of this papers are summarized as follows.

(1) We devise a quadratic polynomial smooth function and change the double regularization support vector machine into an unconstrained optimization function. At the same time, we use a positive function to replace the non-differentiable $\|w\|_1$ approximately. The model becomes an unconstrained and differentiable problem.

(2) We use the steepest descent algorithm to solve the optimization model.

(3) There are three main parameters which decide the experiment effect of the double regularization model. We use the chaotic particle swarm optimization to select the optimal parameters.

This paper is organized as follows. In Section 2, we briefly introduce the L2-norm SVM and L1-norm SVM with some loss functions, then describe our Improved Doubly Regularized Support Vector Machine (IDRSVM) with a loss function in section 3. In section 4, the Chaotic Particle Swarm Optimization (CPSO) algorithm is briefly introduced and is used to optimize the parameter selection of IDRSVM in section 5. The last two sections are the numerical experiments that is demonstrated the effectiveness of our method and the conclusions.

2 L2-svm and L1-svm with the loss function

Given a training dataset $\{(x_i, y_i)\}_{i=1}^n$, where x_i is a vector with p predictor variables and $y_i \in \{-1, 1\}$ denotes the class label. Based on the principle of Structural Risk Minimization (SRM) and the nature of Statistical Learning Theory (SLT) [Vapnik, 2013], some researchers proposed the following L2-norm Support Vector Machine(L2-SVM), which can be seen in Figure 2.

$$\begin{cases} \min\limits_{(w,b)} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^n \xi_i \\ subject\ \ to\ \ y_i((w\cdot x_i)+b) \geq 1-\xi_i \end{cases} \tag{5}$$

Here ξ is a slack variable and this kind of SVM was widely used in data analysis. Some researchers consider that L2-SVM can be equivalently transformed into the "loss+penalty" forma.

$$\min_{(w,b)} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n}[1 - y_i((w\cdot x_i)+b)]_+ \tag{6}$$

The last item in the model (6), x_+ is a positive function and the optimization problem becomes an unconstraint optimization. The L2-norm penalty achieves the bias-variance tradeoff and reduces the variance of the estimated coefficients. It can bring the better prediction accuracy. But the model only emphasises the group effect in the "large p, small n" problem. Mostly this model can not automatically select variables. As far as the L1-norm penalty is concerned, it tends to select only one or few of the variables, especially for highly correlated and relevant variables [Pelckmans et al., 2005]. If we replace the L2-norm with the L1-norm, the unconstraint optimization problem can become the following format:

$$\min_{(w,b)} \|w\|_1 + C\sum_{i=1}^{n}[1 - y_i((w\cdot x_i)+b)]_+ \tag{7}$$

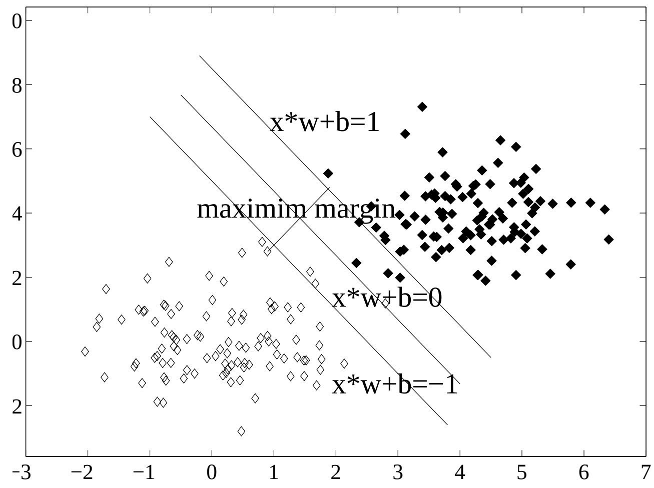

Figure 2: L2-norm support vector machine. The white $\diamond$ denote the positive and black $\diamond$ denotes the negative which are divided into two classes by SVM.

This is an L1-norm Support Vector Machine (L1-SVM) without any constraint condition. With the help of L1-norm and the positive function, this model can be used to perform automatic variable selection. Being similar to the L2-norm penalty, the L1-norm penalty can also improve the prediction accuracy of classification and reduce the coefficients of irrelevant variables exactly. The parameter C can balance the regularization term and the estimated item. But the L1-norm function is non-differentiable and it can brings some trouble in the process of performing data analysis [Wu and Zhang, 2011]. The next section will introduce the algorithm with the help of some smooth functions.

3 IDRSVM

3.1 A smooth function

In recent years, some researchers proposed many kinds of smooth functions which have the different characters. One of the polynomial smooth functions can be expressed as the following equation [Deng et al., 2012]:

$$L(x,k) = \begin{cases} x & : \ x > \frac{1}{k} \\ \frac{(kx+1)^2}{4k} & : \ -\frac{1}{k} \leq x \leq \frac{1}{k} \\ 0 & : \ x < -\frac{1}{k} \end{cases} \tag{8}$$

This differentiable smooth function is smoother and more accurate than that of the positive function x_+ near the zero point. [Yuan et al., 2005] gave the

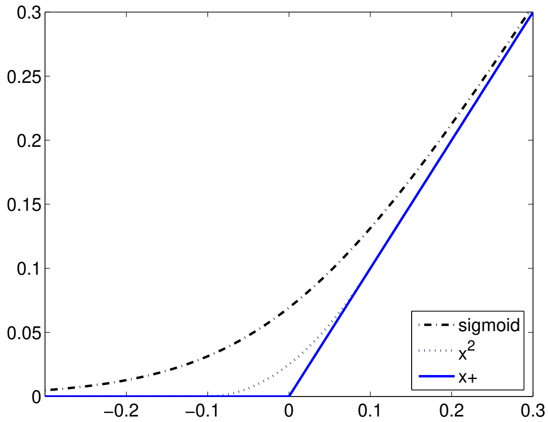

Figure 3: The compare of the three loss functions at the inflection point which is reflected around the zero point.

difference with the plus function in his paper. If a smooth function is defined as the formula (8) and x_+ is the positive function, for any given x and k, the following lemma is true.
(1) $L(x,k) \geq x_+$
(2) $L(x,k)^2 - x_+^2 \leq \frac{1}{19k^2}$

From the lemma, we can see the polynomial smooth function has higher precision than that of the function x_+ under the same k value. Even though compared to the sigmoid function: $L(x,k) = x + \frac{1}{k}log(1+e^{-kx}),\ k > 0$, the formula (8) has better accuracy at the elbow. The compare about the three loss functions can be found around zero point in Figure 3.

From the lemma and Figure 3, we can know that the differentiable quadratic polynomial smooth function can be taken place of the x_+ in a certain extent [Mustafa et al., 2011]. On the other hand, the positive function x_+ is equivalent to the function $max\{0,x\}$ and it is a non-differentiable at the zero point. So we can have the following equation.

$$\max\{0,x\} = x_+ \approx L(x,k) \tag{9}$$

Here k is a given number, when $k = 10$, and the error is about 0.0005. We can gain the relations $max\{0,x\} = x_+ \approx L(x,k)$ in a certain range. On the other hand, we can have the following equations.

$$\|x\| = 2max\{0,x\} - x \approx 2L(x,k) - x \tag{10}$$

So we can use the quadratic polynomial smooth function to substitute the L1-SVM approximately, which can bring some convenient conditions for the model, when we solve the optimization problem. Of course, we must select the proper k and not decrease the accuracy of the model (4).

3.2 IDRSVM

Basing on the advantages of the L1-SVM and the L2-SVM, we can gain the Doubly Regularized Support Vector Machine (DRSVM) in the following expression.

$$\begin{cases} \min\limits_{(w,b)} \frac{C_2}{2}\|w\|_2^2 + C_1\|w\|_1 + \sum_{i=1}^n \xi_i \\ s.\ t.\ y_i((w \cdot x_i) + b) \geq 1 - \xi_i \\ \xi_i \geq 0, i = 1, 2, 3, ...n. \end{cases} \tag{11}$$

C_2, C_1 are the balance parameters. With the help of the quadratic polynomial smooth function and the positive function, we get the approximate deformation of formulation (12).

$$\min_{(w,b)} \frac{C_2}{2}\|w\|_2^2 + 2C_1 L(w, k) - C_1 w + \Sigma_i^n L(y_i(w \cdot x_i), b, k) \tag{12}$$

Where $L(x, k)$ is the quadratic polynomial smooth function and there are three parameters in our model (the smooth factor k and the trade-off parameter C_1, C_2). A researcher [Yuan et al., 2005] gave parameter k the upper bound in the optimization problem.

$$k_{p_2}(n, \varepsilon) \leq \sqrt{\frac{0.0909n}{2\varepsilon}} \tag{13}$$

Where n is the sample number and ε is the accuracy of the smooth function. The formula (12) is an unconstrained optimization problem and it is the first order continuous differentiable. The solutions of this model are difficult to find from the dual optimization problem. An exact line search is used to solve the formula (12) easily [Wardi et al., 2015].

The steepest descent algorithm for the improved DRSVM
(1)Set $X^0 = (w^0, b^0)$ and $\varepsilon > 0$, let k⇐0.
(2)Compute $\nabla f(X^k)$, if $\nabla f(X^k) < \varepsilon$, stop, take $X^k = (w^k, b^k)$, otherwise go to(3)
(3)Compute $p_k = -\nabla f(X^k)$
(4)Seek t_k and make: $f(X_k + t_k p_k) = \min\limits_{t>0} f(X_k + tp_k)$let: $X^{k+1} = X^k + tp_k, k = k + 1$,then go to(2)

4 CPSO

4.1 Chaotic mapping and PSO

The Chaotic Optimization Algorithm (COA) [Wei, 2009] is a recently proposed population-based stochastic optimization algorithm. With the help of some chaotic maps, it can select the better points as the current optimum points. The chaotic ergodicity, regularity, initial sensitivity and topological transitivity are used during the process of optimization. COA is a stochastic search method that differs from any of the existing swarm intelligence optimization methods [Assarzadeh and Naghsh-Nilchi, 2015]. Their several chaotic sequences can be selected in the algorithm and the logistic maps are frequently used chaotic behavior maps and chaotic sequences. In this paper, the logistic maps can be expressed as the following equation [Eberhart and Kennedy, 1995].

$$Cr_{(t+1)} = k \times Cr_{(t)} \times (1 - Cr_{(t)}) \tag{14}$$

Where control variable ($k \in [0,4]$) is the parameter of the logistic mapping which is in the chaotic state and generates chaotic sequences in (0,1). Generated sequences are not periodic and converge, but it must converge to one specific value outside the given range. On the other hand, the standard real-binary Particle Swarm Optimization (PSO) algorithm is a search algorithm based on simulating the social behavior of birds within a flock. This algorithm is proposed by [Eberhart and Kennedy, 1995]. In this algorithm, the velocity V and the position X of each particle will be changed according to the following expressions:

$$V_{id}^{t+1} = wV_{id}^{t} + c_1 r_1 (P_{id}^{t} - X_{id}^{t}) + c_2 r_2 (P_{gd}^{t} - X_{id}^{t}) \tag{15}$$

$$X_{id}^{t+1} = X_{id}^{t} + V_{id}^{t+1} \tag{16}$$

Where w is the inertia weight to be employed to control the impact of the velocity of previous history. V_{id} is the i^{th} particle velocity at iteration d^{th}. Generally, the value of the velocity in V can be clamped to the range $[-V_{max}, V_{max}]$ for controlling excessive roaming of the particle outside the search space. X_{id} is the i^{th} current particle position at iteration d^{th}. r_1, r_2 are the random number between (0,1). c_1, c_2 are the learning or acceleration factors. In the PSO algorithm, the maximal generations or the best position of the particle may be the stopping criteria [Eberhart and Kennedy, 1995]. So the PSO algorithm has shown its robustness and efficacy in solving complex optimization problems.

4.2 CPSO

Compared with many other metaheuristic algorithms, PSO algorithm has several advantages (simple mathematical model and relative simple possibility of

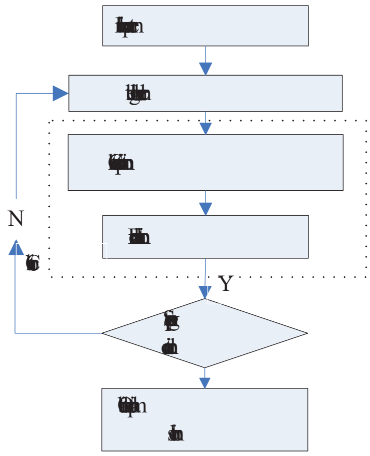

Figure 4: The flow of the Chaotic Particle Swarm Optimization

implementation) [Wei, 2009]. But each coin has two sides, there are some disadvantages in the PSO algorithm. It is prone to premature convergence (especially when dealing with the complex multimodal search problems) and the ability of local optimization is poor. One of the ways to overcome rapid convergence is embedding the Chaotic Optimization Algorithm (COA). COA with the ergodicity, randomicity and regularity is regarded as an optimizer to enhance optimization performance of the PSO algorithm [Wei, 2009]. This algorithm keeps the legality and diversity of solution in one population. Under the chaotic logistic map, the variable traverses all the states in a certain range in order to obtain the optimal solution. To enhance the global optimal solution of the refined search, COA is introduced into the PSO algorithm (namely the CPSO algorithm). It easily jumps out of the local minimum. In the CPSO algorithm, the chaotic scrambling thoughts are respectively added into the initial position and optimum position of particle, which enhances the quick searching ability of the PSO algorithm in the initial stage. The chaotic idea can help the algorithm to jump out of the local extreme values and achieve global optimum. The flow of the CPSO is described as Figure 4.

5 CPSO+IDRSVM algorithm

5.1 Evaluation criterions

For their small sample, non–linear, local minimum and high dimension, the standard L2-SVM is a widely used tool for classification problems. L1-SVM is a variant of the standard L2-SVM, which constrains the L1-norm of the fitted coefficients. Two models have different emphasises in the process of data analysis. The

L2-norm penalty of L2-SVM is to help groups of correlated variables to be selected together. This model tends to make highly correlated input variables, which is often called the grouping effect. Different from the L2-SVM, the L1-SVM has the property of automatically selecting variables. But when there are several highly correlated variables, the L1-SVM tends to pick only a few of them, and removes the rest, furthermore, the number of selected variables of the L1-SVM is upper bounded by the size of the training data [Ghorbanzad'e and Hossein, 2012]. Based on the L1-SVM and the L2-SVM, a Doubly Regularized Support Vector Machine (DRSVM) is proposed to fit the data analysis with high dimensions and small samples. There are several parameters (regularization parameter C_1, C_2, smooth coefficient k) in this model which have a great effect on the performance in the practical application. So we will use the CPSO algorithm to select the parameters effectively. Classification accuracy is used to design a fitness function which is defined as follows:

$$Fitness_i = \frac{correct\ number\ of\ objects}{number\ of\ objects} \tag{17}$$

To fully characterize the classifier performance, a confusion matrix is considered to assess the credibility of the classifier which can be shown in Table 1. The capital letters T, F, P and N express true, false, positive and negative respectively.

	positive	negative
positive	TP	FP
negative	FN	TN%

Table 1: A confusion matrix

It is very important for the imbalanced data classification, where even a total error in predicting a rare class [Huang et al., 2013]. It would have only a small impact on the total accuracy. The specificity and the sensitivity can be defined as the following.

$$Specificity = \frac{TN}{TN + FP} \tag{18}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{19}$$

$$Accuracy = \frac{TP + TN}{TN + PN + FP + FN} \tag{20}$$

Apart from the upper evaluation criteria, a kind of the Matthews's correlation coefficient (MCC) [Mustafa et al., 2011] is also defined to characterize the classier performance with imbalanced class distribution.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \quad (21)$$

Where TP and TN are the number of the true positive and true negative, respectively. FP and FN are the number of the false positive and false negative. Accuracy is used as a statistical measure of how well a binary classification test correctly identifies or excludes a condition. It is the proportion of true results (both true positives and true negatives) among the total number of cases examined. MCC can reflect both sensitivity and specificity of the prediction algorithm.

5.2 Algorithm

Under the help of the smooth function and the plus function, we gain the model (12) which approximates to model (11) in a certain error range. The optimization steps of the CPSO+IDRSVM method are described in the following table:

6 Experiments and discussions

6.1 Descriptions of datasets

In order to verify the effectiveness of our proposed method, five pattern recognition problems with different feature dimensions are used to show the performance of the classifier. These datasets are obtained from the UCI machine learning repository [Asuncion and Newman, 2007]. A description of the datasets are given here:

(1) Abalone dataset

Abalone dataset often is used to predict the age of abalone from physical measurements. After some preprocessing (cutting the shell through the cone, staining it, counting the number of rings through a microscope) the age of abalone can be found. This dataset contains 4177 instances and 8 attributes. The negative number dividing the positive number (imbalance ratio) is 40.

(2) Yeast dataset

Yeast dataset is another dataset form the UCI machine learning repository, which is used to predict the cellular localization sites of proteins. It contains 1484 instances and 9 features (8 predictive, 1 name).

(3) Haberman dataset

Haberman's survival data contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients. All of them had undergone surgery for breast cancer. There are 306 instances and 4 attributes.

(4) Wisconsin Diagnostic Breast Cancer (WDBC)

The optimization steps of CPSO+IDRSVM method
Step1:Initialize parameters (population size N, maximum number of iteration T_{max}, current iteration t=1, learning factor c_1, c_2, inertia weight $[w_{min}, w_{max}]$, velocity range $[v_{min}, v_{max}]$, position range $[x_{min}, x_{max}]$).
Step 2: Select the fitness function (17) which is used to evaluate the performance of the IDRSVM method and calculate the fitness value of each particle.
Step 3: Initialize a vector $Cr_i^d(0)(d = 1, 2, 3, ..., D)$ and generate chaotic queues $Cr_i^d(t)$ by the logistic map (14).
Step 4: Transform the chaotic queues into the range of parameter of IDRSVM according to $X_i^d(t) = X_{min}^d + (X_{max}^d - X_{min}^d)Cr_i^d(t)$.
Step 5: Run the steepest descent algorithm about the IDRSVM model and calculate the fitness function of accuracy.
Step 6: Obtain the individual best P_{ibest}^d and global G_{ibest}^d and judge the stopping criteria (a sufficiently good fitness value or maximum iteration). Go to step 10.
Step 7: Update V_i and X_i of each particle. At the same time c_1, c_2, r_1, r_2 and w are obtained.
Step 8: Compare the fitness value of each particle with its individual best P_{ibest}^d and global G_{ibest}^d then update them as current position and velocity.
Step 9: Determine the end condition. If the end condition is met, the searching process is ended and return to the result of the current best individual. Otherwise, return to Step 5 to recalculate until the termination condition is met or the number of iteration T_{max} is achieved.
Step 10: The obtained optimal position is the values of parameters of the IDRSVM model.
Step 11: Obtain the optimized CPSO+IDRSVM model.

Breast cancer is the one of the popular current cancer in UCI datasets. It is the second largest cause of cancer deaths among women. The WDBC dataset contains 569 instances and 32 features and has 30 inputs that are continuous and classify a tumor as either benign or malignant.

(5) Wisconsin Breast Cancer Dataset (WBCD)

Wisconsin breast cancer dataset was created by Wolberg from the University of Wisconsin. It contains 699 instances and 9 features. In this dataset, 241 instances are malignant and 458 instances are benign.

Dataset	sample	features	positive	negative	imb.ratio
Abalone	4177	8	103	4074	40
Yeast	1484	8	51	1433	28
Haberman	306	3	81	225	2.5
WDBC	569	10	241	458	1.9
WBCD	699	32	212	357	1.7

Table 2: Details of the datasets: the biggest ratio is 40 and two smaller ratio are 1.9 and 1.7 respectively.

6.2 Data analysis

The evaluation criterion includes sensitivity, specificity, the overall classification accuracy, Matthews correlation coefficient and running time. CPSO+IDRSVM model will be performed on the five different imbalance ratio datasets. At the same time, the experiment results will compare with the SVM and the DRSVM which is reported in other journals.Our experiments are coded and executed on the same computer in MATLAB 7.12. Table 3 and Table 4 present the results corresponding to the five different imbalance ratio datasets (Abalone, Yeast, Haberman, WDBC and WBCD), respectively. These datasets have the different imbalance ratio from 40 to 1.7. In all the datasets, the performance metrics of the 10 runs are averaged and reported. Sensitivity, specificity, accuracy and MCC are shown in two tables. These values demonstrate the ability of the proposed classifier. By comparison with other mentioned classifiers, the results of the two tables show that testing accuracy and MCC of the CPSO+IDRSVM classifier are better than other classifiers in every three datasets.

Dataset	Method	SE	SP	MCC	ACC	Time(s)
Abanole	SVM	5.23	97.54	15.45	85.61	4.47
	DRSVM	7.43	89.63	19.98	79.66	8.13
	CPSOIDRSVM	**5.93**	**86.43**	**14.87**	**82.84**	**11.37**
Yeast	SVM	39.05	97.89	43.42	82.64	3.63
	DRSVM	38.90	96.78	44.32	86.34	6.04
	CPSOIDRSVM	**40.31**	**95.82**	**42.78**	**81.21**	**12.67**
Haberman	SVM	20.08	90.13	56.45	85.27	2.13
	DRSVM	21.15	89.69	58.12	82.04	5.36
	CPSOIDRSVM	**22.34**	**89.91**	**60.23**	**86.35**	**19.45**

Table 3: The classification results for the bigger imbalance ratio datasets. The evaluation criterions are the sensitivity (SE), specificity (SP), classification accuracy (ACC), MCC and running time.

Dataset	Method	Se	Sp	Mcc	Acc	Time(s)
Haberman	SVM	20.08	90.13	56.45	85.27	2.13
	DRSVM	21.15	89.69	58.12	82.04	5.36
	CPSOIDRSVM	**22.34**	**89.91**	**60.23**	**86.35**	**9.45**
WDBC	SVM	95.20	97.80	94.45	95.08	2.03
	DRSVM	95.43	96.63	96.18	94.66	4.35
	CPSOIDRSVM	**95.93**	**94.43**	**96.64**	**95.25**	**10.56**
WBCD	SVM	97.70	99.40	91.42	93.64	2.43
	DRSVM	91.90	96.78	94.32	97.34	5.89
	CPSOIDRSVM	**93.31**	**95.02**	**94.78**	**97.81**	**14.78**

Table 4: The classification results for the smaller imbalance ratio data sets as the second part and the assessment indicators and methods are same to Table 3

From results of the tables, the following points can be seen: the imbalance ratio of datasets in Table 3 is bigger than that of in Table 4. For the three datasets in Table 4, the experiment effect of the Haberman datasets is obvious and the CPSO+IDRSVM classifier is the best classifier with 86.35 means testing accuracy. The other two classifiers are 85.27 and 82.04, respectively. For the Abalone and Yeast, those kinds of results cannot be seen from the table. But the running time of Haberman is the longest. We think it is connected with the imbalance ratios and the CPSO algorithm. This seems to be more obvious from Table 2 which imbalance ratios are smaller than those of in Table 3. In Table 4, the imbalance ratios of the three datasets are nearly to 2. The testing accuracies of the CPSO+IDRSVM classifiers are better than the other two. In the WDBC dataset, the testing accuracy of the CPSO+IDRSVM classifier is 95.25 and those of the other two algorithms are 95.08 and 94.66, repetitively. In the WBCD dataset, the testing accuracies of the models, in turn, are 93.64, 97.34, 97.81. The CPSO+IDRSVM classifier gets better experimental results in Table 4 than those of in Table 3. It shows that the small imbalance ratio datasets is suitable for the CPSO+IDRSVM classifier. But the disadvantage of this model maybe needs more time to run the working procedure.

7 Conclusion

Because of the different features of the different norms, the L1-SVM and the L2-SVM have different advantages in the process of the high dimensional data analysises. An IDRSVM has been proposed to deal with a different imbalance ratio datasets. At the same time, the CPSO algorithm is introduced to select the parameters which is brought by the improved model. Chaotic sequences overcome the premature convergence and enhance the optimization performance of the PSO. Effectiveness and powerfulness of the CPSO as a global search metaheuristic algorithm, especially in high dimensional spaces, motivate us to design a swarm intelligence based-classifier. Due to this, the CPSO+IDRSVM is used to obtain the decision hyperplanes in the feature space. For the small imbalance ratio datasets, the experiments results show that the performance of the CPSO+IDRSVM classifier is better than those of the SVM classifier and DRSVM classifier. CPSO selecting the parameter of the IDRSVM is very effective for the classification problems. We find the IDRSVM model can deal with the small imbalance ratio datasets effectively. Our results also show that the proposed classifier works well for medical datasets recognition. In these cases, feature selection helps to reduce the amount of unnecessary, irrelevant and redundant features in datasets and improves the classification accuracy with less computational efforts.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (No. 61562001), High Level Scientific Research Cultivation Foundation of Henan University of Science and Technology (No. 2015GJB010) and Ningxia College Scientific Research Project (NGY20140131).

References

[Asuncion and Newman, 2007] Asuncion, A., Newman, D.: "UCI Machine Learning Repository"; Irvine, CA: University of California, School of Information and Computer Science. http://www.ics.uci.edu/mlearn/MLRepository.html

[Assarzadeh and Naghsh-Nilchi, 2015] Assarzadeh, Z., Naghsh-Nilchi, AR.: "Chaotic Particle Swarm Optimization with Mutation for Classification"; Journal of Medicine Signals and Sensors, 05, 1(2015), 12-20.

[Deng and Tian, 2004] Deng, N., Tian, Y.: "New method for datamining SVM"; Beijing, Science Publish Company (2004).

[Deng et al., 2012] Deng, W., Chen, R., Gao, J., Song, Y., Xu, J.: "A novel parallel hybrid intelligence optimization algorithm for a function approximation problem"; Computer Mathes Apply, 63, 1(2012), 325-336.

[Eberhart and Kennedy, 1995] Eberhart, R., Kennedy, J.: "A new optimizer using particle swarm theory"; In Proceeding of the Sixth International Symposium on Micro Machine and Human Science, Nagoya, Japan (1995), 39-43.

[Ghorbanzad'e and Hossein, 2012] Mehdi Ghorbanzad'e, Mohammad Hossein Fatemi: "A classification of central nervous system agents by least squares support vector machine based on their structural descriptors: a comparative study"; Chemometrics and Intelligent Laboratory Systems, 110, 1(2012), 102-107.

[Huang et al., 2013] Huang, X., Cao, D., Xua, Q., Shen, L., Huang, J., Liang, Y.: "A novel tree kernel support vector machine classifier for modeling the relationship between bioactivity and molecular descriptors"; Chemometrics and Intelligent Laboratory Systems, 120, 15(2013), 71-76.

[Li and Jia, 2010] Li, J., Jia, Y.: "An improverd elastic net for cancer classification and gene selection"; Acta Automatica Sinia, 36(2010), 976-981.

[Li et al., 2016] Li, J., Wang, Y., Cao, Y., Xu, C.: "Weighted doubly regularized support vector machine and its application to microarray classification with noise"; Neurocomputing, 173(2016), 595-605.

[Li et al., 2006] Li, W., Ji, Z., Hui, Z.: "The doubly regularized support vector machine"; Statistica Sinica, 16, 02(2006), 589-615.

[Li et al., 2008] Li, W., Ji, Z., Hui, Z.: "Hybrid huberized support vector machines for microarray classification"; Bioinformatics, 24, 3(2008), 412-419.

[Mustafa et al., 2011] Mustafa, M., Sulaiman, M., Shareef, H., Khalid, S., Rahim, S., Aliman, O.: "An application of genetic algorithm and least squares support vector machine for tracing the transmission loss in deregulated power system"; Power Engineering and Optimization Conference, IEEE, Shah Alam, Selangor, Malaysia (2011), 375-380.

[Mohabatkar et al., 2011] Mohabatkar, H., Mohammad Beigi,M., Esmaeili, A.: "Prediction of GABAA receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine"; Journal of Theoretical Biology, 281, 1(2011), 18-23.

[Pelckmans et al., 2005] Pelckmans, K., Suykens, J., Moor, B.: "Building sparse representations and structure determination on LSSVM substrates"; Neurocomputing, 64(2005), 137-159.

[Shawe-Taylor and Cristianini, 2004] Shawe-Taylor, J., Christianini, N.: "Kernel Methods for Pattern analysis"; Cambridge University Press, Cambridge, UK (2004).

[Vapnik, 2013] Vapnik, N.: "The Nature of Statistical Learning Theory (Second Edition)"; Springer, NewYork (2013).

[Wu and Zhang, 2011] Wu, Z., Zhang, X.: "Elastic multiple kernel learning"; Acta Automatica Sinica, 37, 6(2011), 693-699.

[Wei, 2009] Wei, C.: "Chaotic particle swarm optimization algorithm in a support vector regression electric load forecasting model"; Energy Conversion and Management, 50, 1(2009), 105-117.

[Wardi et al., 2015] Wardi, Y., Egerstedt, M., Hale, M.: "Switched-mode systems: gradient-descent algorithms with Armijo step sizes"; Discrete Event Dynamic Systems, 25, 4(2015), 571-599.

[Yuan et al., 2005] Yuan, Y., Yan, J., Xu, C.: "Polynomial Smooth Support Vector Machine"; Chinese Journal of Computers, 28, 1(2005), 9-17.