Inf Technol Manag (2013) 14:193–204
DOI 10.1007/s10799-013-0161-1

A loan default discrimination model using cost-sensitive support vector machine improved by PSO

Jie Cao · Hongke Lu · Weiwei Wang ·
Jian Wang

Published online: 17 April 2013

Abstract This study proposes a novel PSO–CS-SVM model that hybridizes the particle swarm optimization (PSO) and cost sensitive support vector machine (CS-SVM) to deal with the problem of unbalanced data classification and asymmetry misclassification cost in loan default discrimination problem. Cost sensitive learning is applied to the standard SVM by integrating misclassification cost of each sample into standard SVM and PSO is employed for parameter determination of the CS-SVM. Meantime, the financial data are discretized by using the self-organizing mapping neural network. And the evaluation indices are reduced without information loss by genetic algorithm for decreasing the complexity of the model. The effectiveness of integrated model of CS-SVM and PSO is verified by three experiments comparing with traditional CS-SVM, PSO–SVM, SVM and BP neural network through real loan default data of companies in China. The corresponding results indicate that the accuracy rate, hit rate, covering rate and lift coefficient are improved dramatically by the developed approach. The proposed method can control the different types of errors distribution with various cost of misclassification accurately, reduce the total misclassification cost largely, and distinguish the loan default problems effectively.

Keywords Particle swarm optimization · Cost sensitive learning · Support vector machine · Loan default prediction · Attribute reduction

J. Cao (✉) · W. Wang
School of Economics and Management, Nanjing University of Information Science and Technology, Nanjing 210044, China
e-mail: cj@nuist.edu.cn

H. Lu
School of Economics and Management, Southeast University, Nanjing 210096, China
e-mail: luhkseu@gmail.com

J. Wang
Jiangsu Jinnong Information Co., Ltd., Nanjing 210019, China

1 Introduction

There is no doubt that loan default discrimination is becoming a critical part of a financial institution's loan approval decision process. Since the 1980s, the national banks and investors have faced more and more serious financial risks due to the increasing financial market volatility. Particularly, this economic recession, initiated in the United States and sweeping the globe, has revealed that financial operation pattern has fundamentally changed and the world financial regulatory system has suffered from an unprecedented challenge. Over the last decade, a number of international top banks have developed sophisticated systems in an attempt to model the credit risk arising from important aspects of their business lines. These models are intended to help banks in quantifying, aggregating and managing risk across geographical and product lines. The outputs of these models also play major roles in banks' risk management and performance measurement processes, including performance-based compensation, customer profitability analysis, risk-based pricing and, to a lesser (but growing) extend, active portfolio management and capital structure decisions [1]. Measure and assessment of the enterprise credit default probability are the key element of the Internal Ratings-Based Approach (IRB) of new Basel Accord. And also that is one of the main input variables of the credit risk assessment model. Especially for China's commercial banks in a transitive and incomplete market economy, the

National Credit Management System has not been established yet, and the assessment of loan default situation is even more urgent and important [2]. Nowadays, exploring a more scientific and effective default discrimination method is a major topic of credit risk research [3]. Meanwhile, the international financial community and the academic study mainly regard the loan default discrimination problems as a classification problem of pattern recognition. According to both financial and non-financial situation of the company, the researchers summed up classification rules and established the credit default discrimination model.

Numerous methods have been put forward to constitute a satisfactory loan default discrimination model to study the credit risk problem. Especially, artificial intelligence and machine learning techniques [e.g. artificial neural networks (ANN), decision trees (DT), support vector machines (SVM), etc.] have been used to solve the loan default discrimination problem. Min, Lee, and Han [4] have proposed a hybrid intelligence approach by integrating genetic algorithms and SVM to enhance the accuracy of loan default prediction. Min and Jeong [5] propose a new binary classification method for predicting corporate failure based on a genetic algorithm, and to validate its prediction power through empirical analysis.Hung and Chen propose a selective ensemble of three classifiers, i.e. the decision tree, the back propagation neural network and SVM. Yang, You and Ji [6] present a novel method combined by the partial least squares (PLS) based feature selection with SVM to study the loan default discrimination problem. According to the previous studies, they showed that machine learning techniques are superior to traditional methods in dealing with loan default discrimination and credit score problems, especially in nonlinear pattern classification [7, 8]. However, the main drawback of these methods is that unbalanced data classification problem is not considered. The traditional approach for designing classifier generally pursues more highly accuracy based on the assumption that all misclassifications have the same cost and the sample number of each class is approximately equal. However, the assumption is invalid in some real applications such as fraud detection, medical diagnosis and loan default discrimination. Two types of errors, rejecting the true and accepting the false, are inevitable in the field of loan default discrimination. The two kinds of cost, the cost of predicting a ‘bad credit’ company as a ‘good credit’ company and the cost of predicting a ‘good credit’ company as a ‘bad credit’ company are different. Compared with the latter, the former will involve banks into a greater risk. Therefore, classical classification algorithms without the consideration of different misclassification cost do not perform well.

Nowadays, a great many solutions to the classification of imbalanced problem have been proposed. These solutions include many different forms of resampling techniques, adjusting the costs of misclassification, adjusting the decision threshold, and recognition-based learning [9]. The most direct method for dealing with the highly imbalanced classification problem is to use cost-sensitive learning [10]. Cost-sensitive mining designs classifier based on the principle of the minimizing the overall errors so that it can better meet the misclassification costs of different situations. A cost-sensitive learner can accept cost information from a user and assign different costs to different types of misclassification errors [11]. Maloof [12] argued that learning from imbalanced data sets and learning when misclassification costs are unequal and unknown can be handled in a similar manner. Zhou and Liu [13] also suggest that cost-sensitive learning is the good method to deal with the two-class imbalanced problem.

Support Vector Machine, one of the new techniques for pattern classification, has been extensively and successfully applied to a variety of domains [14], including credit scoring [15], pattern recognition [16], bioinformatics [17], text classification [18] and image retrieval [19]. However, when faced with imbalanced datasets where the number of negative instances far outnumbers the positive instances, the performance of SVM drops significantly. The approach to dealing with imbalanced datasets using SVM biases the algorithm so that the learned hyperplane is further away from the positive class. This is done in order to compensate for the skew associated with imbalanced datasets which pushes the hyperplane closer to the positive class. This biasing can be accomplished in various ways. Wu and Chang [20] proposed an algorithm that changes the kernel function to develop this bias, while in the kernel matrix is adjusted to fit the training data [21]. Veropoulos et al. [22] suggest the using of different penalty constants for different classes of data, making errors on positive instances costlier than errors on negative instances. The penalty-regularized model deserves much more attention because its straight forward idea gives the model intrinsic coherence with its original prototype of SVM. Indeed, this remedy has broadly been applied and extended to deal with applications. In this paper, cost-sensitive SVM is employed to study the loan default discrimination problem.

As we know, proper parameters setting can improve the SVM classification accuracy. To design a SVM, one must choose a kernel function in obtaining the optimal solution. The kernel functions used most frequently are the polynomial, sigmoid and radial basis kernel function (RBF). The RBF is generally applied most frequently, because it can classify multi-dimensional data, unlike a linear kernel function. Additionally, the RBF has fewer parameters than a polynomial kernel. RBF and other kernel functions have similar overall performance. Consequently, RBF is an effective option for kernel function. As a result, this study applies an RBF kernel function in the SVM to obtain optimal solution. Two major RBF parameters in SVM including C and σ must be set appropriately. Parameter C represents the cost of the penalty and the choice of value for C influencing on the classification outcome. If C is too big, the classification accuracy rate will be very high in the training phase, but very low in the testing phase. If C is too small, the classification accuracy rate will be unsatisfactory. Parameter σ has a much greater influence on classification outcomes than C, because its value affects the partitioning outcome in the feature space. An excessively big value for parameter σ results in over-fitting, while a disproportionately small value leads to under-fitting [23]. A growing number of techniques are employed in SVM to improve the possibility of a correct choice of parameter values. Pai and Hong [24] have proposed an SA-based approach to obtain parameter values for SVM. Huang, Chen and Wang [25] adopted the GA to optimize the feature subset and model parameter selection for the SVM. Lin et al. [26] proposed a novel PSO–SVM model that hybridizes PSO and SVM to improve the classification accuracy with the SVM kernel parameter setting.

Besides two parameters C and σ, other factors, such as the quality of the evaluation indices, may influence the classification accuracy rate. Datasets with unimportant, noisy or highly correlated features will significantly decrease the classification accuracy rate. By removing these features, the efficiency and classification accuracy rate can be obtained. Approaches for feature selection can be categorized into two models, namely a filter model (e.g. factor analysis, independent component analysis and discriminate analysis) and a wrapper model (e.g. GA and PSO) [26]. In the wrapper model, Zhang, Jack, and Nandi [27] develop a GA-based approach to discover a beneficial subset of features for SVM in machine condition monitoring. Huang and Wang [28] present a GA-based feature selection and parameters optimization for SVM. Huang and Dun [29] employ PSO to select appropriate input feature subset. In this paper, GA is employed to reduce the evaluation indices without information loss. Cao and Lu [30] apply the improved PSO to determine the parameters of multiclass LS-SVM for improving loan default classification accuracy.

This study introduces a new technology which named PSO–CS-SVM in loan default discrimination. The PSO is used for parameters selection of the CS-SVM. Further more, we use attribute reduction which decrease the training samples dimensions, so that the classifier will be more efficient. Compared with other algorithms, our supposed PSO–CS-SVM has a better improvement in accuracy rate, hit rate, covering rate and lift coefficient. The data test result proves that the method can control the different types of error distribution with various cost of misclassification, reduce the total misclassification cost, and distinguish the customer value.

This paper is organized as follows. Section 2 introduces basic SVM and cost-sensitive SVM. Section 3 describes the improved PSO and PSO-based CS-SVM. In the Sect. 4, attribute reduction based on GA is proposed. Section 5 presents the new Loan Default Discrimination Prediction model called PSO–CS-SVM. In Sect. 6 presents the experimental results from using the proposed method to classify the real loan default datasets of companies in China. Section 7 gives remarks and draws a conclusion.

2 Cost-sensitive support vector machine

2.1 Basic concepts of the SVM classifier

In this section, the basic SVM concepts for typical two-class classification problems are described briefly [31]. Given a training set of instance-label pairs (x_i, y_i), $i = 1, \ldots, n$, where $x_i \in R^l$, $y_i \in \{+1, -1\}$. SVM finds an optimal separating hyperplane with the maximum margin by solving the following optimization problem:

$$\begin{aligned} &\text{Min} && \tfrac{1}{2} w^T w \\ &\text{subject to} && y_i(x_i \cdot w + b) - 1 \geq 0 \end{aligned} \tag{1}$$

It is known that to solve this quadratic optimization problem we must find the saddle point of the Lagrange function:

$$L_p(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} (\alpha_i y_i (x_i \cdot w + b) - 1) \tag{2}$$

where the $\alpha_i \geq 0$ denotes Lagrange multipliers. It is necessary to search for an optimal saddle, because $L_p(w, b, \alpha)$ must be minimized with respect to the primal variables w and b and maximized with respect to the non-negative dual variable α_i. By differentiating with respect to w and b, and introducing the Karush Kuhn-Tucker (KKT) condition for the optimum constrained function, $L_p(w, b, \alpha)$ can be transformed to the dual Lagrange equation $L_D(\alpha)$:

$$\begin{aligned} &\text{Max} && L_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \tfrac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \\ &\text{subject to} && \alpha_i \geq 0 \; i = 1, \ldots, n \text{ and } \sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned} \tag{3}$$

To find the optimal hyperplane, a dual Lagrange equation $L_D(\alpha)$ must be maximized with respect to non-negative α_i. The solution α_i for the dual optimization problem determines the parameters w^* and b^* of the optimal hyperplane. Thus the optimal hyperplane decision function $f(x) = \text{sgn}(w^* \cdot x + b^*)$ can be expressed as:

$$f(x) = \mathrm{sgn}\left(\sum_{1}^{n} y_i \alpha_i^* (x_i \cdot x) + b^*\right) \quad (4)$$

In a typical classification task, only a small subset of the Lagrange multipliers α_i usually tends to be greater than zero. Geometrically, these vectors are the closest to the optimal hyperplane. The respective training vectors having nonzero α_i are called support vectors, as the optimal decision hyperplane $f(x, a^*, b^*)$ depends on them exclusively [30].

Next, the non-separable case (no-linear SVM) can be extended based on the above concepts. In terms of these introduced slack variables, the problem of finding the hyperplane that provides the minimum number of training errors has the formal expression as follows:

$$R(w,\xi) = \frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{n}\xi_i\right)$$
$$\text{subject to} \quad y_i(x_i \cdot w + b) + \xi_i - 1 \geq 0, i = 1, \ldots, n, \ \xi_i \geq 0 \quad (5)$$

Where $C \geq 0$ is a penalty parameter on the training error, and ξ_i is the non-negative slack variable.

This optimization model can be solved using the Lagrangian method, which is almost equivalent to the method for solving the optimization problem in the separable case. One must maximize the same dual variables Lagrange equation $L_D(\alpha)$ (Eq. (6)) as in the separable case.

$$L_D(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j)$$
$$\text{subject to} \quad 0 \leq \alpha_i \leq C, i = 1, \ldots, n \text{ and } \sum_{i=1}^{n}\alpha_i y_i = 0 \quad (6)$$

To find the optimal hyperplane, a dual Lagrange equation $L_D(\alpha)$ must be maximized with respect to non-negative α_i under the constrains $\sum_{i=1}^{n}\alpha_i y_i = 0$ and $0 \leq \alpha_i \leq C$, $i = 1, \ldots, n$. And the penalty parameter C is the upper bound on α_i.

2.2 Cost-sensitive support vector machine

The above formulations implicitly penalize errors in both classes equally. However, as described in the introduction, there may be different costs associated with the two different kinds of errors. To address this issue, cost-sensitive extensions of the SVM have been proposed. Data sets with different class distributions lead to the effect that conventional machine learning methods are biased towards the larger class. To overcome this problem and obtain sensitive and accurate classifiers, we extend and improve the standard SVM formulation. However, the problem is that with imbalanced datasets, the learned boundary is too close to the positive instances. We need to bias SVM in a way that will push the boundary away from the positive instances. Veropoulos et al. [22] suggest using different error costs for the positive (c_+) and negative (c_-) classes.

Given a training set of instance-label pairs (x_i, y_i, c_i), $i = 1, \ldots, n$, where $x_i \in R^l$, $y_i \in \{+1, -1\}$and c_i denotes misclassification cost of the sample x_i. The SVM finds an optimal separating hyperplane with the maximum margin by solving the following optimization problem:

$$\min R(w,\xi) = \frac{1}{2}\|w\|^2 + C\sum_{\{i|y_i=+1\}}^{n} c_+\xi_i + C\sum_{\{i|y_i=-1\}}^{n} c_-\xi_i$$
$$\text{subject to} \quad y_i(x_i \cdot w + b) \geq 1 - \xi_i, \ \xi_i \geq 0, i = 1, \ldots, n \quad (7)$$

where c_+ and c_- denote the misclassification cost of false positive and false negative. C is a penalty parameter on the training error and ξ_i is the non-negative slack variable.

Different with the function (5), in the Function (7), misclassification cost is taken into account. To solve the optimization problem (7), the following Lagrange equation is constructed.

$$L_p = \frac{1}{2}w \cdot w + C\sum_{\{i|y_i=+1\}}^{n} c_+\xi_i + C\sum_{\{i|y_i=-1\}}^{n} c_-\xi_i - \sum_{i=1}^{n}\alpha_i\{y_i(x_i \cdot w + b) - 1 + \xi_i\} - \sum_{i=1}^{n}\beta_i\xi_i \quad (8)$$

where α_i denotes Lagrange multipliers, hence $\alpha_i \geq 0$. The search for an optimal saddle point is necessary because L_p must be minimized with respect to the primal variables w, b, ξ_i and maximized with respect to the non-negative dual variable α_i. By differentiating with respect to w, b, ξ_i, the following equations are obtained:

$$\frac{\partial L_p}{\partial w} = w - \sum_{i=1}^{n}\alpha_i y_i x_i = 0$$
$$\frac{\partial L_p}{\partial b} = -\sum_{i=1}^{n}\alpha_i y_i = 0 \quad (9)$$
$$\frac{\partial L_p}{\partial \xi_i} = co_i C - \alpha_i - \beta_i = 0$$

The expression of the dual Lagrange represents below:

$$W = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{n}\alpha_i$$
$$\text{subject to} \quad \textstyle\sum_{i=1}^{n}\alpha_i y_i = 0 \quad (10)$$
$$0 \leq \alpha_i \leq Cc_+, \quad \forall i : y_i = +1$$
$$0 \leq \alpha_i \leq Cc_-, \quad \forall i : y_i = -1$$

3 The improved algorithm

3.1 The improved particle swarm optimization

The particle swarm optimization (PSO) is a computation intelligence technique, which was motivated by the organisms' behavior such as schooling of fish and flocking of birds. PSO can solve a variety of difficult optimization problems. One advantage is that PSO makes the most of the physical movements of the individuals in the swarm and has a flexible and well-balanced mechanism to enhance the global and local exploration abilities. The other advantage is its simplicity in coding and consistency in performance. The global optimizing model proposed by Shi and Eberhart [32] is described as follows:

$$v_{k+1} = w \cdot v_k + c_1 r_1 (pbest_k - x_k) + c_2 r_2 (gbest_k - x_k)$$
$$x_{k+1} = x_k + v_{k+1} \tag{11}$$

where v_k is the velocity of particle i, represents the distance to be travelled by this particle from its current position; k is the number of iterations; x_k represents the particle position; w is the inertial weight; c_1 and c_2 are the positive constant parameters; r_1 and r_2 are the random functions in the range [0, 1]; $pbest_k$ (local best solution) is the best position of the kth particle and $gbest_k$ (global best solution) is the best position among all particles in the swarm.

In general, the parameters w, c_1, c_2, r_1, r_2 are the important factors which influence the convergence of the PSO. However, parameters r_1 and r_2 cannot guarantee ergodicity of the optimization entirely in phase space because they are absolutely random in the traditional PSO. By studying the basic model above, many researches have shown that if the acceleration Constant c_1 and c_2, as well as parameters such as maximum speed are too large, particle swarm may miss the optimal solution, resulting in algorithm does not converge; In the case of convergence, all the particles move in the direction of its best solution and all the particles tend to the same state, making the latter part of the convergence rate significantly slows. Moreover, while algorithm converges to a certain accuracy, it cannot continue to be optimized any more. Thus the last accuracy of the algorithm is not very high.

To solve the problem, the convergence factor χ and inertia weight ω can be applied to improve the basic particle swarm optimization model. Xia and Dong propose a new improved form by synthesized the existing model of PSO. To search for the optimal solution, each particle changes its velocity according to the cognition and social parts as follows [33]:

$$v_{k+1} = \lambda \cdot [\omega v_k + c_1 r_1 (pbest_k - x_k) + c_2 r_2 (gbest_k - x_k)] \tag{12}$$

Where c_1 indicates the cognition learning factor; c_2 indicates the social learning factor, and r_1 and r_2 are random numbers uniformly distributed in $U(0, 1)$.

$$\lambda = \frac{2}{\left|2 - c - \sqrt{c^2 - 4c}\right|} \tag{13}$$

$$\omega = \omega_{\max} - \frac{(\omega_{\max} - \omega_{\min}) \cdot T_{\max}}{T} \tag{14}$$

Each particle then moves to a new potential solution based on the following equation:

$$x_{k+1} = x_k + v_{k+1} \tag{15}$$

3.2 Parameter determination

As the penalty function C and the kernel function σ will affect the performance of CS-SVM, merely different impact on different data sets. Therefore, the parameter selection is a key issue on the successful application of the algorithm. In this paper, our improved PSO algorithm is used for parameters selection. To implement our proposed approach, this research used the RBF kernel function [defined by Eq. (16)] for the SVM classifier because the RBF kernel function can analyze higher dimensional data. The formulation can be given as below:.

$$K(x, x_k) = \exp(-\|x - x_k\|^2 / 2\sigma^2) \tag{16}$$

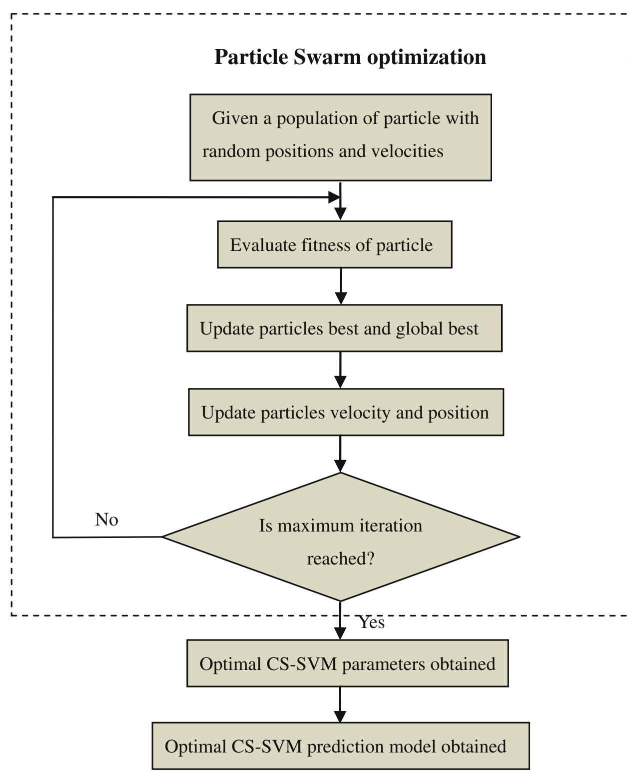

Fig. 1 The process of optimizing the CS-SVM parameters with PSO

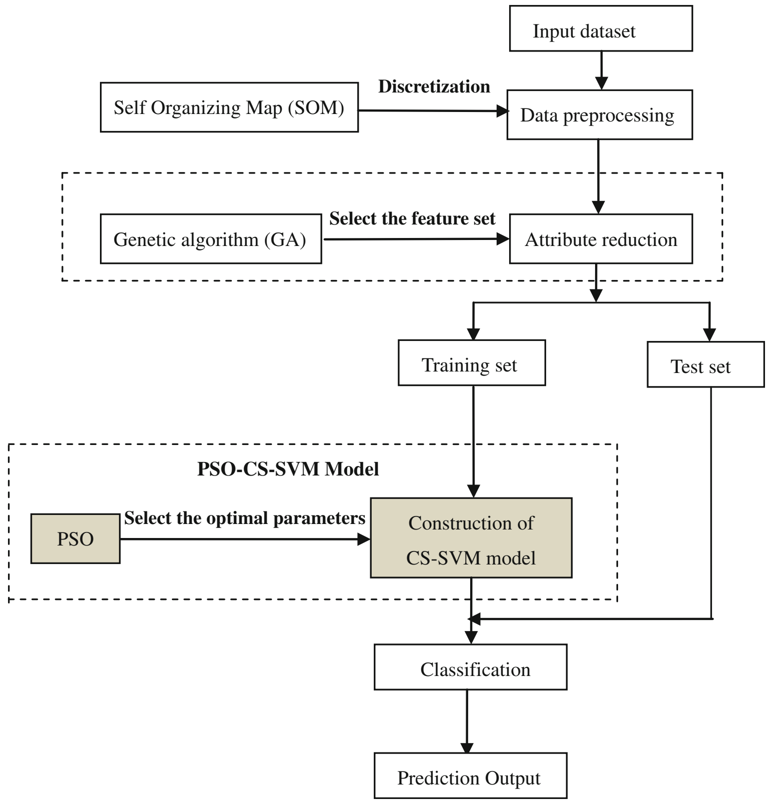

Fig. 2 The proposed PSO–CS-SVM model for loan default discrimination

Additionally, selecting a suitable fitness function plays an important role in evolutionary algorithms. In this paper, mean absolute percentage error (MAPE) is used for the fitness function which can be defined as below.

$$f = 1/n \sum_{1}^{n} |(y_i - \hat{y}_i)/2y_i| \tag{17}$$

Where n stands as the number of the test sample, y_i is the actual value, $\hat{y}_i$ is a predicted value and f is the value of the fitness. Using improved particle swarm algorithm, the parameter of selection process of CS-SVM is as follows, shown in Fig. 1:

The concrete steps are shown below:

Step 1: (Initialization) Initialize the parameters of particle swarm optimization algorithm. In a D-dimension space, randomly generate D-dimensional position $x_{i1}^t, x_{i2}^t, \ldots, x_{in}^t$ for the particle i at iteration t and constitute the initial population $X(t)$.and also randomly generate the velocity of each particle $v_{i1}^t, v_{i2}^t, \ldots, v_{in}^t$ and build up the velocity matrix $V(t)$.

Step 2: (Fitness) Measure the fitness of each particle in the population and the formulation can be defined as below:

$$f = 1/n \sum_{1}^{n} |(y_i - \hat{y}_i)/2y_i|$$

Step 3: (Comparison) For each particle, Compare the current fitness $f(x_i)$ with the best position fitness $f(pbest_k)$ of the group, if $f(x_i) < f(gbest_k)$, then global optimal solution $gbest_k = x_i$.

Step 4: (Update) Compute the velocity of each particle with Eq. (12) and Eq. (15), Generate new population $X(t + 1)$. Speed adjustment rules are as follows

$$v_i = \begin{cases} V_{\max}, & v_i > V_{\max} \\ -V_{\max}, & v_i < -V_{\max} \end{cases} \tag{18}$$

Step 5: The iteration is terminated if the number of iteration reaches the pre-determined maximum number of iteration and returns the current best individual as a result; Return to Step 2 otherwise.

Step 6: (Output). Through the above processes, the optimal parameters of C and σ in the cost-sensitive support vector machines are easily gotten.

4 Attribute reduction

Datasets with unimportant, noisy or highly correlated features will significantly decrease the classification accuracy rate. By removing these features, the efficiency and classification accuracy rate can be obtained. Attribute reduction is one of the most important topics in knowledge discovery; however, it is a NP-Hard problem to select an optimal subset from attribute set. Attribute reduction refers to the ability to maintain the same conditions and to delete irrelevant or redundant knowledge without losing any information. Existing reduction algorithms, mainly from the core of rough set, employ heuristic search method to exploit minimum reduction. In the rough set theory, $\beta = pos_C(D_j) - pos_{C-\{c_i\}}(D_j)$ is employed to examine the importance of condition attribute $c_i \in C$ corresponding to decision attribute D_j. If $\beta = 0$, condition attribute c_i is considered to be redundant and can be deleted.

In this paper, a genetic algorithm is implemented to find out the minimum reduction [34]. There in, a set of evaluation indices are defined as $C = \{c_1, c_2, \ldots, c_n\}$, Ω denotes the evaluation space. Chromosome is an *n-bit* binary string, each binary bit corresponding to an evaluation index. If the value of the binary bit is equal to 1, we select its corresponding evaluation index, otherwise, we remove it. In this way, each chromosome corresponds to an attribute subset of Ω. The algorithm's fitness function f is defined below [35]:

$$f(\theta) = 1 - \frac{L_\theta}{n} + \frac{C_\theta}{(m^2 - m)/2} \tag{19}$$

Where L_θ denotes the number of evaluation indices chosen from chromosome θ; C_θ denotes the number of rows covered by chromosome θ in decision table; f denotes the number of evaluation indices; m denotes training sample.

5 The loan default discrimination model of PSO-CS-SVM

In this paper, the hybrid improved model is formulated to deal with the loan default discrimination problem by incorporating the advantages of the PSO and CS-SVM. The proposed hybrid loan default discrimination model is composed of 6 parts, shown in Fig. 2

The algorithm of our loan default discrimination model is as follows:

Step 1: Data pre-processing and normalization; normalization processing of data is crucial with the goal of speeding up the convergence of model and reducing the impact of imbalance of the data capacity to the classifier. Linear differential analysis is used which can be defined as $x'_{ij} = \frac{x_{ij} - \min_i}{\max_i - \min_i} \in [0, 1]$, where $\max_i$, $\min_i$ denote maximum and minimal value of all sample data in attribute C_i respectively; x_{ij} denotes the *i-th* attribute in the *j-th* sample; x'_{ij} denotes the data after being normalized.

Step 2: Discretization. Since the rough set algorithm can only deal with the discrete attribute, continuous data should be discretized first. One of the most important ANN is the Self-Organization Map (SOM) proposed by Kohonen. In this network there are an input layer and the Kohonen layer which are usually designed as two-dimensional arrangement of neurons that maps n-dimensional input to two dimensional. It is a competitive network with the characteristic of self-organization providing a topology-preserving mapping from the input space to the clusters [36, 37]. In this paper, SOM neural network is used for discretization. If the number of clusters is specified, discrete results can reflect the data distribution objectively. Usually, the Euclidean distance is used to compare each node with each object although any other metric could be chosen. The Euclidean distance between an object with observed vector $\vec{P}_k$ and the weight vector $\vec{W}_j$ is given by

$$d_j = \left[\sum_{i=1}^{N} (\vec{p}_i^k - \vec{w}_{ji})^2 \right]^{\frac{1}{2}} \tag{20}$$

The Kohonen update rule for the winner node is given by

$$w_{ji}(t+1) = w_{ji}(t) + \eta(t)[p_i^k - w_{ji}(t)] \tag{21}$$

Where $j = 1, 2, \ldots, M$ and $\eta(t)$ is learning rate.

Step 3: Attribute Reduction; Evaluation indices are reduced by GA without information loss and then features subset can be obtained.

Step 4: Construction of CS-SVM model. In the process of the model, the trained sample which has been

Table 1 The loan default discrimination financial indexes

Factors	The evaluation indices
Ability to pay	Current ratio (C_1); quick ratio (C_2); cash ratio (C_3); current assets to total debt ratio (C_4); asset-liability ratio (C_5); interest cover (C_6)
Financial efficiency	Return on net worth (C_7); return on total assets (C_8); primary business profit margin (C_9); net profit margin (C_{10}); ratio of profits to cost (C_{11});
Working capital position	Accounts receivable turnover (C_{12}); current asset turnover (C_{13}); inventory turnover (C_{14}); turnover of total capital (C_{15});
Ability development	Growth rate of main business income (C_{16}); net profit growth rate (C_{17}); total asset growth rate (C_{18}); net asset growth rate (C_{19});

discretisized is fed into CS-SVM. Meanwhile, the proposed improved PSO algorithm is used to determine the two parameters of CS-SVM.

Step 5: Classification. The proposed PSO–CS-SVM model that has been trained is employed to determine the category for the test sample. The out of classification is the prediction categories for each test record.

Step 6: According to the dataset which is randomly selected, CS-SVM, PSO-SVM, SVM and BP neural network are used to compare to our proposed PSO–CS-SVM model.

6 Empirical analysis

6.1 The evaluation indices selection and data descriptions

In order to verify the proposed approaches in loan default prediction, 372 Chinese A-share has listed companies of manufacturing industry in Shanghai and Shenzhen Stock Market are selected as sample data. The specific financial data of 372 listed companies are collected from the annual financial reports of these companies on the financial website called Hexun.com. Sample interval is selected as 2005–2007. According to the evaluation system taken by domestic banks, rating agencies and financial research [3, 38, 39], the paper sets up an index system made up of 19 financial indicators, such as current ratio, quick ratio, cash ratio and other financial information which synthetically reflect to the financial condition of listed companies. Table 1 shows the features/variables used in datasets. Therein, 320 companies are 'non-ST' companies whose financial position is sound with a low default risk. These companies are considered as 'good' companies and $y = 1$. The remained 52 companies are 'ST or *ST' companies whose financial position is not so good with a high default risk is. These companies are referred to as 'bad' companies and $y = -1$ [40]. Meanwhile, if the 'bad' company is misclassified as 'good' company, we call it as the type I error. On the contrary, if the 'good' company is misclassified as 'bad' company, it is recorded as type II error. For the purpose of enhancing the generalization ability and discrimination accuracy of the new model, a random sampling method is employed to divide the data set into two parts randomly. The whole 240 companies, composed of 208 'good' companies and 32 'bad' companies, are selected as a training sample set, while the remaining 132 companies as a test sample set. The test sample set composed of 112 positive samples and 20 negative samples are taken to testify the prediction accuracy of the proposed PSO–CS-SVM. The experiments are carried out 3 times.

6.2 Data reprocessing and attribute reduction

In this paper, the 19 evaluation indices are discretized by using the SOM neural network. Next, the evaluation indices are reduced without information loss by a genetic algorithm. By means of genetic algorithm, one minimum reduction including 7 evaluation indices is achieved below: $\{C_2, C_6, C_7, C_{11}, C_{14}, C_{17}, C_{19}\}$. Compared with the key evaluation indices which are used by current domestic and international scholars to study loan default discrimination problem, the reduced 7 evaluation indices can reflect to a company's credit position and subordinated loan capacity.

6.3 System implementation details

Our implementation platform is carried out on the Matlab7.1, a mathematical development environment, by extending the Libsvm version 2.82 which was originally designed by Chang and Lin [41]. The empirical evaluation is performed on AMD Turion 64 X2 Duo Core CPU running at 2.0 GHz and 2 G MB RAM. And the formulation of the RBF kernel function formulation is $K(x, x_k) = \exp(-\|x - x_k\|^2/2\sigma^2)$. In the model of PSO–CS-SVM, the number of the population size is 20 and the maximum number of iterations is 100. We set $c_1 = c_2 = 1.3$, $\lambda = 0.729$, $\omega_{\max} = 0.9$, $\omega_{\min} = 0.1$, and initial inertia weight $\omega = 0.9$.According to the misclassification cost value, if the 'bad' company is misclassified to 'good' company, the misclassification cost is $c_- = 5$; If the 'good' company is misclassified to 'bad' company, the misclassification cost is $c_+ = 2$, in other condition $c_i = 0$.

In our experiment, we compare the performance of our classifier with four other popular methods: (I) CS-SVM, (II) PSO-SVM, (III) classical SVM, (IV) BP. In the CS-SVM and classical SVM algorithms, we set $\sigma = 2.1$ and $C = 7$.

6.4 Comparison with other models

The performance of the loan default discrimination is evaluated by a confusion matrix as illustrated in Table 2. Therein, In this matrix, TN (true negative) represents 'bad' company correctly classified, FP (false positive) represents 'good' company incorrectly classified as 'bad' company, FN (false negative) represents 'bad' company incorrectly classified as 'good' company, and TP (true positive) represents the 'good' company correctly classified. A perfect forecast program would have values in cells 'TN' and 'TP' only. In the real world, imperfect forecasts would result in the values of cells 'FN' and 'FP'. The misclassification cost values can be given by do-main experts, or learned via other approaches. In this paper, the cost of false negative (denoted as FN) is 5 and the cost of the false positive 2. Based on the confusion matrix, several measures can be computed, including Expected loss function and the average misclassification cost. Also accuracy rate, hit rate, covering rate and lift coefficient can be computed. The formulation of above performance indices are listed after the confusion matrix displayed as below:

The error rate of 'bad' company classified as 'good' company is $F_{A,B} = FN/(TN + FN)$. The error rate of 'good' company classified as 'bad' company is $F_{B,A} = FP/(FP + TP)$. The accuracy rate can be defined as $T = (TN + TP)/(TN + FP + FN + TP)$. Expected loss function [42] is defined as $C = FP \times c_{+} + FN \times c_{-}$. The average misclassification cost which is calculated by $ave(C) = (FN \times F_{A,B} \times c_{-} + FP \times F_{B,A} \times c_{+})/(TN + FP + FN + TP)$. Hit rate can be represented as $TN/(TN + FP)$; covering rate is $TN/(TN + FN)$; lift coefficient is defined as $TN/(dr * (TN + FP))$. Where dr means actual default rate, the formulation is $(TN + FN)/(TN + FP + FN + TP)$.

Table 2 confusion matrix

	Predict bad (A)	Predict good (B)
Actual bad (A)	TN	FN
Actual good (B)	FP	TP

PSO–CS-SVM which has been trained is used to determine the category for the 132 test sample including 20 negative samples (Sample I) and 112 positive samples (Sample II). Tables 3, 4 and 5 has shown the number of the true negative (TN), false positive (FP), false negative (FN) and true positive (TP), also shown the $F_{A,B}$ and $F_{B,A}$. $F_{A,B}$ means misclassification rate of the Sample I, also means the error rate of 'bad' company classified as 'good' company, while $F_{B,A}$ means misclassification rate of the Sample II, also means the error rate of 'good' company classified as 'bad' company .

Based on results of Tables 3, 4 and 5, we figure out accuracy, hit rate, Covering rate and Lift with different models in three experiments. The results of these four performance indices are given in Table 6 clearly. From Table 6, we can know that for the same model, results of 3 experiments have minor differences. Thus, the average prediction results with different models in detail are shown in Table 7. As Sample I and Sample II are imbalanced in

Table 3 Classification results of the 1st group

Group	Model	TN	FN	Sample I	$F_{A,B}$ (%)	FP	TP	Sample II	$F_{B,A}$ (%)
1	PSO–CS-SVM	10	10	20	50.00	4	108	112	3.57
	CS-SVM	9	11	20	55.00	5	107	112	4.67
	PSO-SVM	8	12	20	60.00	3	109	112	2.68
	SVM	2	18	20	90.00	1	111	112	0.89
	BP	2	18	20	90.00	3	109	112	2.68

Table 4 Classification results of the 2nd group

Group	Model	TN	FN	Sample I	$F_{A,B}$ (%)	FP	TP	Sample II	$F_{B,A}$ (%)
2	PSO–CS-SVM	10	10	20	50.00	3	109	112	2.68
	CS-SVM	7	13	20	65.00	2	110	112	1.79
	PSO-SVM	4	16	20	80.00	2	110	112	1.79
	SVM	2	18	20	90.00	1	111	112	0.89
	BP	1	19	20	95.00	4	108	112	3.57

Table 5 Classification results of the 3rd group

Group	Model	TN	FN	Sample I	$F_{A,B}$ (%)	FP	TP	Sample II	$F_{B,A}$ (%)
3	PSO–CS-SVM	11	9	20	45.00	3	109	112	2.68
	CS-SVM	8	12	20	60.00	2	110	112	1.79
	PSO-SVM	3	17	20	85.00	1	111	112	0.89
	SVM	1	19	20	95.00	1	111	112	0.89
	BP	2	18	20	90.00	4	108	112	3.57

Table 6 Comparison of four performance indices in data mining

Group	Model	Accuracy (%)	Hit rate (%)	Covering rate (%)	Lift
1	PSO–CS-SVM	89.39	71.43	50.00	4.71
	CS-SVM	87.88	64.29	45.00	4.24
	PSO-SVM	88.64	72.73	40.00	4.80
	SVM	85.61	66.67	10.00	4.4
	BP	84.09	40.00	10.00	2.64
2	PSO–CS-SVM	90.15	76.92	50.00	5.08
	CS-SVM	88.64	77.78	35.00	5.13
	PSO-SVM	86.36	66.67	20.00	4.4
	SVM	85.61	66.67	10.00	4.4
	BP	82.58	20.00	5.00	1.32
3	PSO–CS-SVM	90.91	78.57	55.00	5.19
	CS-SVM	89.39	80.00	40.00	5.28
	PSO-SVM	86.36	75.00	15.00	4.95
	SVM	84.85	50.00	5.00	3.3
	BP	83.33	33.33	10.00	2.2

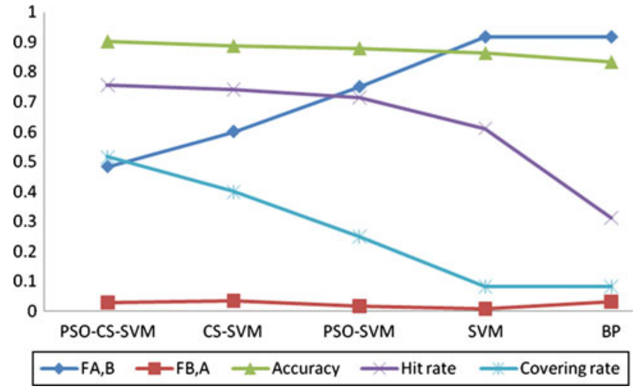

Fig. 3 Comparison among different models

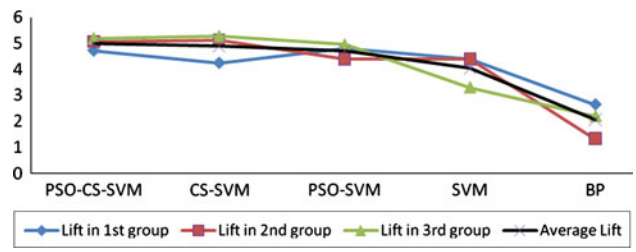

Fig. 4 Comparison of the Lift coefficient output with different models

our three experiments, if the model has a lower error rate of Sample I, it means that the model has a better performance on data imbalanced problem. From Table 7, for PSO–CS-SVM, its average $F_{A,B}$ is 48.30 %,average accuracy rate is 90.15 %, average hit rate is 75.64 %, average covering rate is 51.67 % and average lift coefficient is 4.99. For CS-SVM, its average $F_{A,B}$ is 60.00 %,average accuracy rate is 88.64 %, average hit rate is 74.02 %, average covering rate is 40.00 % and average lift coefficient is 4.88. For PSO-SVM, its average $F_{A,B}$ is 75.00 %,average accuracy rate is 87.88 %, average hit rate is 71.47 %, average covering rate is 25.00 % and average lift coefficient is 4.72. It is obvious that $F_{\mathrm{A,B}}$ of PSO–CS-SVM is lower than other 4 models. It means that PSO–CS-SVM has a better performance on data imbalanced problem. Additionally, the error rate of 'bad' company classified as 'good' company in CS-SVM model is lower than PSO-CSM, classical SVM and BP. Nevertheless, compared with CS-SVM, PSO-SVM, SVM and BP, the error rate of 'good' company classified as 'bad' company is not the lowest. If $F_{\mathrm{B,A}}$ and $F_{\mathrm{A,B}}$ both can be dropped off, the overall prediction will be improved further.

It also can be seen from Figs. 3 and 4 that PSO–CS-SVM has a distinct improvement in accuracy rate, hit rate, covering rate and lift coefficient ($\mathrm{F_{A,B}}$, $\mathrm{F_{B,A}}$). In Fig. 3, the $\mathrm{F_{B,A}}$ of is not the lowest among all models, we can endure this error because this factor means the 'good' company is classified as 'bad', this is a security error in loan default discrimination. FCS-SVM without parameter-selection and PSO-SVM without cost sensitive learning is both inferior to PSO–CS-SVM in these four performance indices.

Next, another two performance indices such as C and $ave(C)$ are focused on. The prediction results of three experiments on these two indices are given in Table 8. For PSO–CS-SVM, average expected misclassification lost is 56, and the average misclassification cost is 17.89 %. For CS-SVM, average expected misclassification lost is 66, and average misclassification cost is 27.57 %. For PSO-SVM, average expected misclassification lost is 79.67, and average misclassification cost is 43.56 %. For SVM, average expected misclassification lost is 93.67, and average misclassification cost is 63.71 %. For BP, average expected misclassification lost is 99, and average misclassification cost is 63.89 %. Vividly as shown in Fig. 5, PSO–CS-SVM using cost-sensitive learning and parameters determination can obtain a better classification performance in expected loss and the average misclassification cost. By means of algorithm analysis, CS-SVM has employed the cost-sensitive learning, however, the parameters of CS-SVM is not the optimal value. Though PSO is employed for parameter determination of SVM, cost-sensitive learning is not considered in the model PSO-SVM. It is clear that cost-sensitive learning is necessary when dealing with the problem of unbalanced data classification, good parameters selection will improve the classification accuracy of the SVM. The empirical results show that PSO–CS-SVM can improve the

Table 7 Average prediction results of four performance indices

Model	$F_{A,B}$ (%)	$F_{B,A}$ (%)	Accuracy (%)	Hit rate (%)	Covering rate (%)	Lift
PSO–CS-SVM	48.30	3.02	90.15	75.64	51.67	4.99
CS-SVM	60.00	3.57	88.64	74.02	40.00	4.88
PSO-SVM	75.00	1.79	87.88	71.47	25.00	4.72
SVM	91.67	0.89	86.27	61.11	8.33	4.03
ANN(BP)	91.67	3.27	83.33	31.11	8.33	2.05

Table 8 Prediction results with different models

Model	1		2		3		Average	
	C	Ave (C) [%]	C	Ave (C) [%]	C	Ave (C) [%]	C	Ave (C) [%]
PSO–CS-SVM	58	19.16	56	19.06	54	15.46	56	17.89
CS-SVM	65	23.25	69	32.14	64	27.33	66	27.57
PSO-SVM	66	27.44	86	48.54	87	54.75	79.67	43.56
SVM	92	61.38	92	61.38	97	68.38	93.67	63.71
ANN(BP)	96	61.49	103	68.59	98	61.58	99	63.89

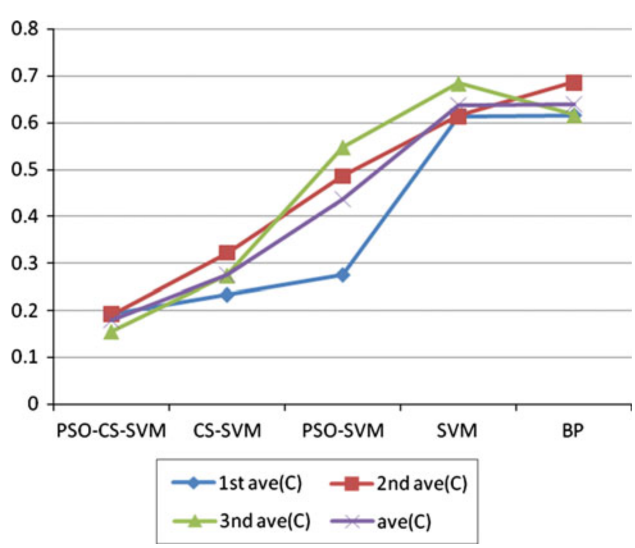

Fig. 5 Comparison of the average prediction with different models

accuracy of loan default discrimination, reduce the loan default misclassification cost and distinguish the loan default problems effectively.

7 Conclusions and expectations

Loan defaults discrimination plays an important role in credit risk assessment and the determination of economic capital. Traditional data mining method for the loan default discrimination problems generally ignores the unbalanced data classification and asymmetry misclassification cost of credit data set. In this paper, cost-sensitive learning method is applied to SVM of different penalty coefficients to construct the loan default discrimination model. PSO algorithm is employed to select the optimal parameters of the CS-SVM. A Genetic algorithm based approach is formulated to obtain a subset of beneficial features. This optimal subset of features is then adopted in both training and testing to obtain the optimal outcomes in classification. Comparison of the obtained results with those of other approaches demonstrates that the developed PSO–CS-SVM approach has a better classification performance than others tested. Especially, the proposed method can control the different types of errors distribution with various cost of misclassification accurately, reduce the total misclassification cost largely, and distinguish the loan default problems effectively. The study of this paper provides a new way to deal with the problem of unbalanced data classification and asymmetry misclassification cost in the binary-class classification problem of loan default discrimination, also improves the classification performance of credit rating. However, this paper only considers the binary-class classification problem of the company credit status. More complex multi-class classification problem of credit rating needs to be further studied. Meanwhile, this study shows experimental results with the RBF kernel. However, other kernel parameters can also be optimized using the same approach. Experimental results obtained from real world loan default datasets, other public datasets can also be tested in the future to verify and extend this approach.

Acknowledgments The authors would like to thank the anonymous referees for their valuable comments and suggestions. Their comments helped to improve the quality of the paper immensely. This work is partially supported by NSFC (60804047), the Science and Technology Project of Jiangsu province, China (BE2010201), Ministry of education, humanities and social sciences research project (11YJCZH005), Jiangsu provincial department of education philosophy and social science project (2010SJB790025) and the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

1. Bhekisipho T (2009) Multiple classifier application to credit risk assessment. Expert Syst Appl 37(4):3326–3336
2. Ma RW, Tang CY (2007) Building up default predicting model based on logistic model and misclassification loss. Syst Eng Theory Pract 27:33–38
3. Ke KL, Feng ZX (2008) Short-term loan default prediction based on integration of rough sets and genetic algorithm. Syst Eng Theory Pract 28(4):27–34
4. Min SH, Lee J, Han I (2006) Hybrid genetic algorithms and support vector machines for bankruptcy prediction. Expert Syst Appl 31(3):652–660
5. Min JH, Jeong C (2009) A binary classification method for bankruptcy prediction. Expert Syst Appl 36:5256–5263
6. Yang ZJ, You WJ, Ji GL (2011) Using partial least squares and support vector machines for bankruptcy prediction. Expert Syst Appl 38:8336–8342
7. Huang Z, Chen H, Hsu CJ et al (2004) Credit rating analysis with support vector machines and neural networks: a market comparative study. Decis Support Syst 37(4):543–558
8. Ong CS, Huang JJ, Tzeng GH (2005) Building credit scoring models using genetic programming. Expert Syst Appl 29(1): 41–47
9. Chawla N, Japkowicz N, Kolcz A (2004) Editorial: special issues on learning from imbalanced data sets. SIGKDD Explor 6:1–6
10. Mccarthy K, Zabar B, Weiss G (2005) Does cost-sensitive learning beat sampling for classifying rare classes? In: proceedings of the ACM SIGKDD first international workshop on utility-based data mining. ACM Press, pp. 69–75
11. Tsai CH, Chang LC, Chiang HC (2009) Forecasting of ozone episode days by cost-sensitive neural network methods. Sci Total Environ 407(6):2124–2135
12. Maloof M (2003) Learning when data sets are imbalanced and when costs are unequal and unknown. In: proceedings of the ICML-2003 workshop: learning with imbalanced data sets II, pp. 73–f80
13. Zhou ZH, Liu XY (2006) Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Trans Knowl Data Eng 18:63–77
14. Li Z, Ling L, Lian D (2012) Business intelligence in enterprise computing environment. Inf Technol Manage 13:297–310
15. Yu L, Yao X, Wang SY, Lai KK (2011) Credit risk evaluation using a weighted least squares SVM classifier with design of experiment for parameter selection. Expert Syst Appl 38: 15392–15399
16. Pontil M, Verri A (1998) Support vector machines for 3D object recognition. IEEE Trans Pattern Anal Mach Intell 20(6):637–646
17. Yu GX, Ostrouchov G, Geist A, et al (2003) An SVM based algorithm for identification of photosynthesis-specific genome features. Second IEEE computer society bioinformatics conference. CA, USA, pp. 235–243
18. Joachims T (1998) Text categorization with SVM: learning with many relevant features. In: proceedings of ECML-98,10th European conference on machine learning, Vol. 1398
19. Tong S, Chang E (2001) Support vector machine active learning for image retrieval. In: proceedings of ACM international conference on multimedia, pp. 107–118
20. Wu G, Chang E (2003) Class-boundary alignment for imbalanced dataset learning. In: ICML 2003 workshop on learning from imbalanced data sets II. Washington, DC
21. Cristianini N, Kandola J, Elisseeff A, et al (2001) On kernel target alignment. In: Advances in neural information processing systems, vol 14, pp 367–373
22. Veropoulos K, Campbell C, Cristianini N (1999) Controlling the sensitivity of support vector machines. In: Dean T (ed) IJCAI: proceedings of international joint conference on artificial intelligence. Morgan Kaufmann, Stockholm, pp 55–60
23. Pardo M, Sberveglieri G (2005) Classification of electronic nose data with support vector machines. Sens Actuators B Chem 107(2):730–737
24. Pai PF, Hong WC (2005) Support vector machines with simulated annealing algorithms in electricity load forecasting. Energy Convers Manage 46(17):2669–2688
25. Huang CL, Chen MC, Wang CJ (2007) Credit scoring with A data mining approach based on support vector machines. Expert Syst Appl 33(4):847–856
26. Lin SW, Ying KC, Chen SC et al (2008) Particle swarm optimization for parameter determination and feature selection of support vector machines. Expert Syst Appl 35(4):1817–1824
27. Zhang L, Jack LB, Nandi AK (2005) Fault detection using genetic programming. Mech Syst Signal Process 19:271–289
28. Huang CL, Wang CJ (2006) A GA-based feature selection and parameters optimization for support vector machines. Expert Syst Appl 31:231–240
29. Huang CL, Dun JF (2008) A distributed PSO–SVM hybrid system with feature selection and parameter optimization. Appl Soft Comput 8:1381–1391
30. Cao J, Lu HK, Wang WW et al (2012) A novel five-category loan-risk evaluation model using multiclass LS-SVM By PSO. Int Journal Inf Technol Decis Mak 11(4):857–874
31. Burgers CJC (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Disc 2:121–167
32. Shi Y, Eberhart RC (1998) A modified particle swarm optimizer. In: proceeding of the IEEE congress on evolutionary computation, pp. 69–73
33. Xia KW, Dong Y, Du HL (2007) Oil layer recognition model of LS-SVM based on improved PSO algorithm. Control Decis 22(12):1385–1389
34. Tao Z, Xu BD, Wang DW et al (2003) Rough set knowledge reduction approach based on GA. Syst Eng 21(4):116–122
35. Ke KL, Feng ZX (2008) Five-category classification of loan risk based on integration of rough sets and neural network system. Control Theory Appl 25(4):759–763
36. Kohonen T (1989) Self-organization and associative memory. Springer-Verlag, New York
37. Kohonen T (1995) Self-organizing maps. Springer, Berlin. Vol.27, No. 2, pp. 278–279
38. Wu DS, Liang L (2004) Research of credit score based on V-fold cross-validation and elman neural networks. Syst Eng Theory Pract 4:92–98
39. Xue F, Ke KL (2008) Five-category evaluation of commercial bank's loan based on integration of rough sets and neural network. Syst Eng Theory Pract 1:40–45
40. Zhang M, Zhou ZF (2009) An evaluation model for credit risk of enterprise based on multi-objective programming and support vector machines. China soft sci mag 04:185–190
41. Chang CC, Lin CJ (2001) LIBSVM: a library for support vector machines. Software available at: http://www.csie.ntu.edu.tw/cjlin/libsvm
42. Oos P, Vanhoof K, Ooghe H (1999) Credit classification: a comparison of logic models and decision. In: proceedings of European conference on machine learning. Chemnitz: [s. n.]