Loan Prediction Model Based on AdaBoost and PSO-SVM

Tao Zhang [a)], Baodian Li [b)]

Department of Informatics, Beijing University of Technology, Beijing, 100124, China.

[a)] Corresponding author: zhangtony8888@qq.com
[b)] libaodian1234@163.com

Abstract. In view of the problem that the slow manual approval of loans and traditional classification algorithms have low recognition rate for a few sample classes, an integrated learning classification model based on PSO optimization support vector machines is proposed. Particle swarm optimization (PSO) is used to optimize the model parameters of SVM classifier. The AdaBoost integrated learning method is used to integrate SVM weak classifiers, and a loan prediction model based on multi-classifier optimization and integration is established. Taking the Lending Clud loan data set as the research object, the loan prediction model was built. The simulation results show that compared with the standard SVM algorithm and the PSO-SVM algorithm, the AdaBoost-PSO-SVM method can effectively improve the accuracy of the classification of a small number of samples, The classification accuracy of the whole sample and the generalization rate, the accuracy of the model applied to loan prediction is obviously better than other models.

Key words: Integrated learning; SVM; PSO; AdaBoost; loan prediction model.

INTRODUCTION

In recent years, with the rapid growth of the personal credit market, both the online emerging internet loans and traditional bank store modes have faced unprecedented challenges [1]. Online business processes are convenient, but new risks such as cheating and hacking continue to emerge. However, the inherent difficulties of manual auditing of offline loans and the lack of basis for making risk-based decisions have always restricted the normal development of bank lending business.

Data show that at the end of June 2017, the total value of non-performing loans of national banks reached 1.64 trillion yuan, an increase of 123.5 billion yuan or 8.2% over the end of 2016; the average non-performing loan ratio was 1.74%, which remained at 13 years although unchanged from the end of last year High since. Financial institutions risk prevention and control pressures continue to increase, banks should be intelligent transformation, the most important thing is to do smart wind control [2].

In view of the development of loan intelligence, scholars at home and abroad have conducted theoretical researches from different perspectives and proposed various methods for loan intelligence. P Danenas and G Garsva SVM-based credit risk domain classifier selection, put forward a particle swarm optimization based optimal linear SVM classifier selection technology [3]. Jiang M Based on Logistic Regression and Posteriori Probability SVM for Housing Loan Evaluation Model, two single models based on Logistic Regression and Posterior Probability Support Vector Machines are proposed, and a non-negative linear combination forecasting for each single model is constructed model [4]. LI Tai-yong Aiming at the problems of low accuracy and poor interpretability of traditional credit evaluation methods, a model of personal credit evaluation using sparse Bayesian learning method is proposed [5]. Lin Guoqiang and others used the social structure of users to construct P2P user default prediction based on complex networ·ks and machine learning [6]. By constructing a FCLC model based on improved Particle Swarm Optimization (PSO) and multi-class Least Squares Support Vector Machines (LS-SVM), Cao Jie's research and solution to the application and implementation of FCLC in China's microfinance banks rely mainly on subjective judgments, It is difficult to control and reduce the risk of loans [7].

Based on the above research, this paper proposes the use of Particle Swarm Optimization (PSO) algorithm to optimize the penalty parameter C and kernel parameters of SVM. The optimized SVM is regarded as the weak classifier of AdaBoost algorithm, and the weight of each weak classifier is calculated. Then The weight of each PSO-SVM classifier is combined into a strong classifier, and a loan prediction model based on AdaBoost-PSO-SVM is established. The validity of the model on the Lending Clud loan dataset is verified by experiment. Rate has a significant effect.

COMMON ALGORITHMS AND PRINCIPLES

Adaboost Algorithm

AdaBoost is an abbreviation of "Adaptive Boosting" in English. Its adaptivity lies in that: the weight of the sample misclassified by a previous basic classifier will increase, and the weight of a correctly classified sample will decrease, And again used to train the next basic classifier [8]. At the same time, a new weak classifier is added in each round of iteration until the final strong classifier is determined by reaching a predetermined small error rate or reaching a pre-specified maximum number of iterations.

Adaboost algorithm can be briefly described as three steps:

First, the weight distribution D_1 of training data is initialized. Assuming there are N training sample data, each training sample is given the same weight at the very beginning: 1 / N, this trains the sample set's initial weight distribution $\mathrm{D}_1(\mathrm{i})$:

$$D_1(i) = (w_1, w_2, \dots\dots w_n) = \left(\frac{1}{N}, \dots\dots, \frac{1}{N}\right) \tag{1}$$

Iterate t=1, 2..., T

step.1 Select a classifier h with the lowest error rate as the t-th base classifier Ht, Select a classifier h with the lowest current error rate as the t-th base classifier Ht, and calculate a weak classifier: X->{-1, 1}. The error in the distribution of the weak classifier is:

$$e_t = P(\mathrm{H}_t(x_i) \neq y_i) = \sum_{i=1}^{N} w_{ti} I(\mathrm{H}_t(x_i) \neq y_i) \tag{2}$$

step.2 Calculate the weight of the weak classifier in the final classifier (weak classifier weight repressentati--on):

step.3 Update training sample weight distribution D_{t+1}:

$$\alpha_t = \frac{1}{2} ln \left(\frac{1-e_t}{e_t}\right) \tag{3}$$

$$D_{t+1} = \frac{D_t(i) exp\left(-\alpha_t y_i H_t(x_i)\right)}{Z_t} \tag{4}$$

Z_t is the normalization constant,$\mathrm{Z}_t = 2\sqrt{\mathrm{e}_t(1-\mathrm{e}_t)}$

Finally, the classifiers are combined according to the classifier weights, that is:

$$\mathrm{f}(x) = \sum_{t=1}^{T} \alpha_t \mathrm{H}_t(x) \tag{5}$$

By using the sign function sign, a strong classifier is:

$$H_{final} = sign(\mathrm{f}(x)) = sign\left(\sum_{t=1}^{T} \alpha_t \mathrm{H}_t(x)\right) \tag{6}$$

Particle Swarm Optimization

Particle Swarm Optimization (PSO) is a group-based intelligent optimization algorithm proposed by Kennedy and Eberhart in 1995. It is a universal global search algorithm designed by simulating preying behavior of birds [9]. Particle swarm optimization (PSO) simulates a bird in a flock by designing a massless particle. The particle has only two properties: velocity V and position X, where velocity represents the speed of movement and position represents the direction of movement. Each particle searches for the optimal solution separately in the search space and records it as the current individual extremum $\mathrm{P_{best}}$ and shares the extremums with other particles in the entire particle swarm to find the optimal individual extremum as the whole The current global optimal solution $\mathrm{G_{best}}$ of the particle swarm, all particles in the particle swarm adjust their speed and position according to the current individual extremal $\mathrm{P_{best}}$ found by themselves and the current global optimal solution $\mathrm{G_{best}}$ shared by the entire particle swarm.

Let m particles in the N-dimensional target search space form a community. The position of the i-th particle in the N-dimensional search space is represented by the vector $\mathrm{x_i = (x_{i1}, x_{i2}, x_{i3} \dots x_{iN})}$; the velocity is represented by the vector $\mathrm{v_i = (v_{i1}, v_{i2}, v_{i3} \dots v_{iN}), i = 1, 2, \dots m}$.

The particle velocity and position update formula is as follows:

$$\begin{gathered} \mathrm{v_{in}^{k+1} = \omega v_{in}^{k} + c_1 r_1 (p_{in}^{k} \text{-} x_{in}^{k}) + c_2 r_2 (p_{gn}^{k} \text{-} x_{in}^{k})} \\ \mathrm{x_{in}^{k+1} = x_{in}^{k} + v_{in}^{k+1}} \end{gathered} \tag{7}$$

Where: i=1, 2 ... m; n = 1, 2 ... N; k is the number of iterations; $\mathrm{c_1}$ and $\mathrm{c_2}$ are learning factors and are nonnegative constants; $\mathrm{r_1}$ and $\mathrm{r_2}$ are random numbers between [0, 1] , ω is the inertia weight, $\mathrm{v_{in}} \in$ [-vmax, vmax], vmax is a constant set by the user; $\mathrm{p_{in}}$ is the optimal position searched by the ith particle; $\mathrm{p_{in}}$ is the optimal position searched by the entire particle swarm. The iteration termination condition is generally chosen as the maximum number of iterations or the optimal position searched by the particle swarm so far to meet the adaptation threshold.

SVM

Support Vector Machine (SVM), first proposed by Corinna Cortes and Vapnik et al in 1995, is a supervised learning model, which is usually used for pattern recognition [10], classification and regression analysis. High-dimensional pattern recognition shows many unique advantages.

There is a training sample set $\mathrm{A = \{(x_i, y_i), i = 1, 2 \dots n\}}$,where $x_{\mathrm{i}} \in \mathrm{R}$, $y_{\mathrm{i}} \in \{1, -1\}$ is the category of x_{i}, Classification is:

$$y(x) = sgn\{w \cdot \emptyset(x) + b\} \tag{8}$$

Where x, w, b, and φ (x) denote input vectors, weight coefficients, offsets, and feature mappings, respectively. This transforms equation (7) into the following optimization problem:

$$\begin{gathered} \min \quad \frac{1}{2} \|\mathrm{w}\|^2 \\ \text{subject to } \mathrm{y_i[(wx_i) + b]\text{-}1 \geq 0 \ (i = 1, \cdots, l)} \end{gathered} \tag{9}$$

Using Lagrange function to transform the above optimization problem into duality problem, which is:

$$\min \ \frac{1}{2} \sum_{\mathrm{i=1}}^{\mathrm{l}} \sum_{\mathrm{j=1}}^{\mathrm{l}} \mathrm{\alpha_i \alpha_j y_i y_j (x_i x_j)} \text{-} \sum_{\mathrm{j=1}}^{\mathrm{l}} \mathrm{\alpha_j} \tag{10}$$

$$\text{subject to } \sum_{i=1}^{l} \alpha_i y_i = 0\,\alpha_i \geq 0, i = 1, \cdots, l$$

Where α_i is the Lagrange multiplier corresponding to each sample, the above quadratic optimization problem is solved to obtain the optimal α, a positive component α_j of α is selected to calculate $b^* = y_i \text{-} \sum_{i=1}^{l} \alpha_i^* y_i (x_i x)$, and the sample corresponding to positive component a4 is called a support vector. The classification decision function introduced is:

$$f(x) = \operatorname{sgn}\{w \cdot \emptyset(x) + b\} = \operatorname{sgn}\left\{\sum_{i=1}^{l} \alpha_i^* y_i (x_i x) + b^*\right\} \tag{11}$$

Adaboost-PSO-SVM Model Design

SVM Parameter Optimization

Because SVM model parameters directly affect the SVM model classification accuracy, it is particularly important to optimize the parameters of SVM. SVM adjustable parameters only punishment parameters C and kernel function parameters σ. In view of the optimization effect of PSO algorithm, the PSO algorithm is used to optimize the parameters of the SVM model, avoiding the blindness of the traditional artificial test to determine the parameters [11].

PSO optimization SVM parameter process shown in Figure 1:

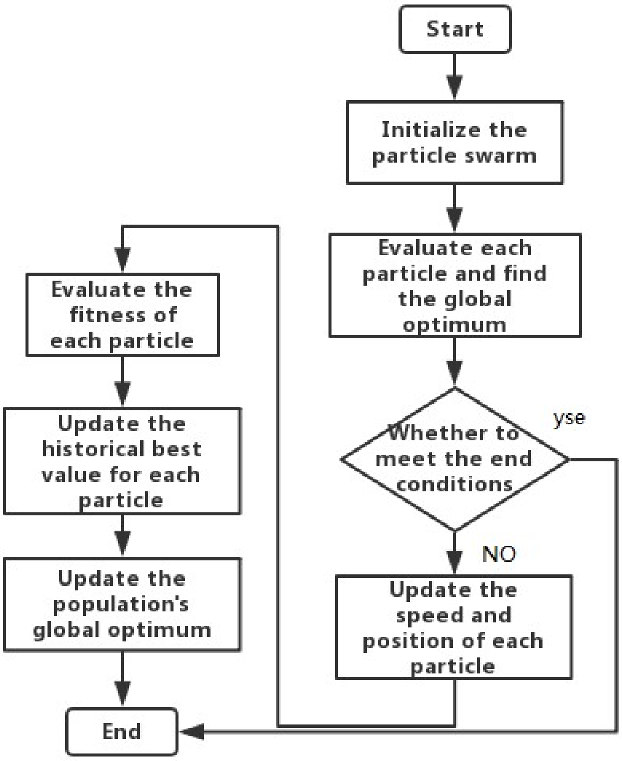
Start
Initialize the particle swarm
Evaluate each particle and find the global optimum
Whether to meet the end conditions
yse
NO
Update the speed and position of each particle
Evaluate the fitness of each particle
Update the historical best value for each particle
Update the population's global optimum
End

FIGURE 1. PSO optimization SVM parameter flow chart.

Adaboost-PSO-SVM Model

Based on the above theory, the PSO-SVM model optimized in the previous section is taken as the base classifier, and then combined with the AdaBoost algorithm for the same sample set to train the PSO-SVM base classifier with better prediction ability. Finally, combined into a strong classifier, to improve the model prediction accuracy and recognition rate of a few samples. Design AdaBoost-PSO-SVM model of the computing process shown in Figure 2.

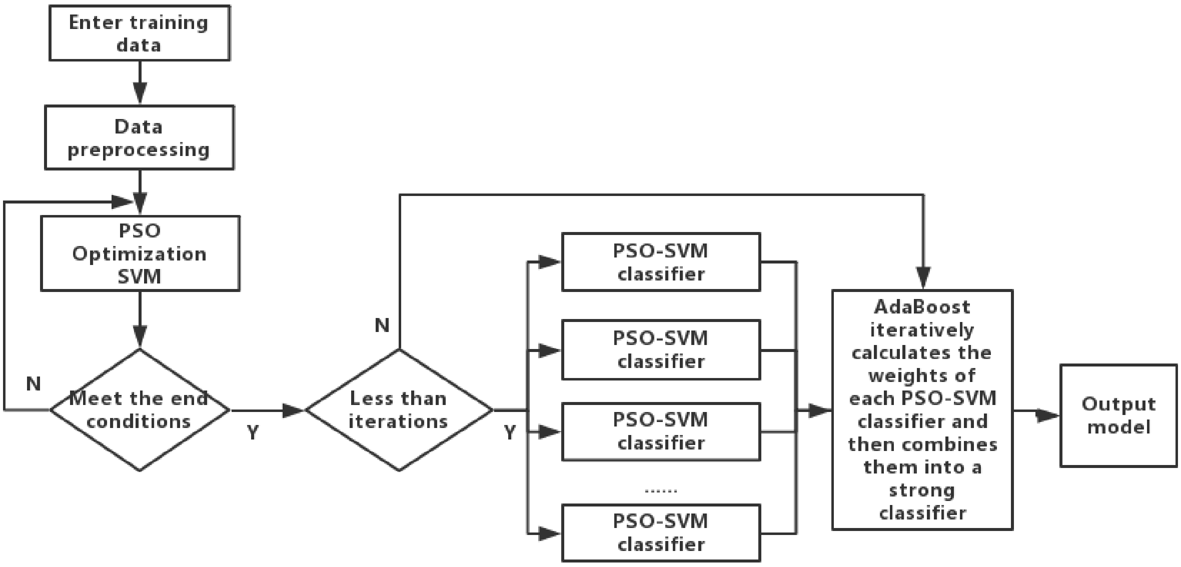

FIGURE 2. AdaBoost-PSO-SVM model operation flow chart.

EXPERIMENTAL RESULTS AND ANALYSIS

SVM Parameter Optimization

The same sample was used in the same search area to search for optimal performance, and the search ability of grid search algorithm, genetic algorithm and particle swarm optimization algorithm in support vector machine penalty parameter C and RBF kernel function parameter σ was compared and analyzed.

As can be seen from Table 1, the optimization results of the three algorithms are very close to those of the SVM. However, the PSO is better than the genetic algorithm and the grid search algorithm when it is used.

TABLE 1. Algorithm optimization performance comparison.

Optimization algorithm	Time / s	Recognition rate/%
GridSearch	56.79	93.54
GA	97.43	94.12
PSO	73.23	95.23

Loan Approval Method Based on Adaboost-PSO-SVM

In order to further verify that AdaBoost method based on PSO-SVM is superior to other intelligent algorithm model loan prediction, 50 experiments on loan prediction model based on SVM, PSO-SVM and AdaBoost-PSO-SVM are carried out respectively. The experimental results the average of 50 experimental results. The commonly used evaluation criteria are Precision and Recall, and F-measure and G-mean values are introduced to compromise Precision and Recall. Table 2 shows the comparison results.

TABLE 2. Comparison of experimental results.

Experimental method	Precision	Recall	F-measure	G-mean
SVM	0.8581	0.8683	0.8631	0.8631
PSO-SVM	0.9034	0.8967	0.9001	0.9000
AdaBoost-PSO-SVM	0.9321	0.9471	0.9395	0.9396

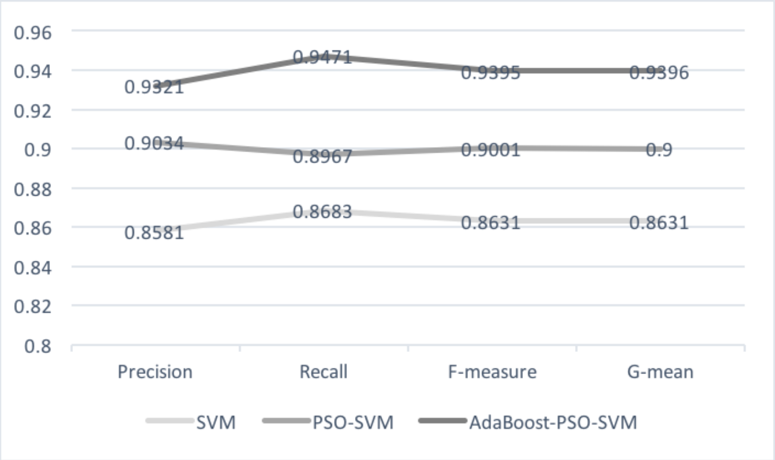

FIGURE 3. Experiment comparison line chart.

As can be clearly seen from Figure 3, the experimental results on the Lending Clud loan dataset using the AdaBoost method based on PSO-SVM are obviously improved compared with the traditional SVM and PSO-SVM. F-measure is a compromise between Precision and Recall. In general, the higher the F-measure value, the better the classifier performance is considered [12]. The G-mean value comprehensively examines the accuracy of the positive predictive classification and the negative categorical classification accuracy, and the G-mean value is high only when both are high. From the F-measure and G-mean data, we can see that the AdaBoost method based on PSO-SVM has a very high classification and prediction ability.

SUMMARY

The research starts from the loan data and analyzes the influence of the data characteristics on the classification accuracy so that the extracted features have good reliability and sensitivity in the actual diagnosis. Not only the feature selection has a great influence on the classification effect of SVM, but also the appropriate parameters of penalty and kernel function can also improve its classification effect. Particle swarm optimization (PSO) is used to search the parameters of SVM and compared with grid search algorithm and genetic algorithm. The experimental results show that the grid search algorithm, the genetic algorithm and the particle swarm optimization are effective in improving the recognition rate in the loan prediction based on SVM. However, in the optimization time, the PSO is better than the grid search Algorithm and Genetic Algorithm.

Finally, considering that the anti-fidelity data set is a set of unbalanced data sets, according to the idea of AdaBoost algorithm, the PSO-SVM is a weak classifier and a strong classifier is formed by weighted combination. In identifying a few types of samples and the overall classification effect has significantly improved.

In summary, the proposed loan approval model not only high classification accuracy, but also practical.

REFERENCES

1. Yao Gao. Research on financial innovation and development strategy of China's commercial banks in Internet finance era [D]. Jilin University of Finance and Economics, 2016.
2. Xingyu Pan. Improving Operational Services and Promoting the Transformational Development and Intelligent Development of Commercial Banks [J]. International Finance, 2012 (8): 32-37.

3. Danenas P, Garsva G. Selection of Support Vector Machines based classifiers for credit risk domain[M]. Pergamon Press, Inc. 2015.
4. Jiang M, Yuan X. Combining evaluation model based on Logistic regression and posterior probability SVM for residential loan[J]. Journal of Natural Science of Heilongjiang University, 2008.
5. Li Tai-yong, Wang Hui-Jun. Personal credit evaluation based on sparse Bayesian learning [J], Journal of Computer Applications, 2013,33(11):3094-3096+3148.
6. Guo-qiang Lin, Yi-ming ZHAO, QUAN Qing-zuo, FAN Ying. Prediction of P2P User Defaults Based on Complex Networks and Machine Learning [J]. Journal of Beijing Normal University (Natural Science) ,2017,53(01):24-27+2.
7. Jie Cao, HongKe Lu. A novel five-category loan-risk evaluation model using multiclass ls-svm by pso [J]. International Journal of Information Technology & Decision Making, 2012, 11(04):857-874.
8. Owusu E, Zhan Y, Mao Q R. A neural-AdaBoost based facial expression recognition system[J]. Expert Systems with Applications, 2014, 41(7):3383-3390.
9. Oliveira L D D, Jeszensky P J E. PSO-based multiuser detectors for high-order modulation DS/CDMA systems under spatial and multipath diversities[J]. International Journal of Wireless & Mobile Computing, 2013, 6(3):221-235.
10. Allauzen C, Cortes C, Mohri M. Large-Scale Training of SVMs with Automata Kernels[C]// International Conference on Implementation and Application of Automata. Springer-Verlag, 2010:17-27.
11. Qin Bo, Wu Qingchao, Juan Juan Zhang.Study on Prediction of Oxygen Demand for BOF Steelmaking Based on PSO Optimization SVM [J]. fluidity & Control Technology, 2014, 33 (12): 121-124.
12. Huaping Guo, Yadong Dong, Haitao Mao.A rare class classification method based on logical discriminant [J]. Microsoft Microcomputer Systems, 2016, 37 (1): 140-145.