ISSN: 2688-8653

Personal Credit Evaluation based on PSO-SVM

Huan Wang, Qingzhi Zheng, Ninghao Yang and Xiufang Yuan *

School of Mathematics and Statistics, Sichuan University of Science & Engineering 644000, China

* yxiufang1990@163.com

Abstract

Support vector machines have more applications in the field of credit evaluation, and the parameters of support vector machines have a greater impact on the classification effect. Therefore, the particle swarm optimization algorithm is used to optimize its parameters, and the experimental data is combined with the information gain method to reduce dimensionality After processing, a good classification effect and accuracy are obtained.

Keywords

support sector machines, information gain ,particle swarm optimization, personal credit evaluation.

1. Introduction

Regarding the issue of personal credit evaluation, the United States has proposed a $FICO$ scoring system to assess the risk of default. The domestic credit system is developing day by day. Liu Xiaoya and others use the $C4.5$ algorithm to optimize support vector machines to evaluate the credit system[1]. Li Jiarong and others used BP neural network to evaluate the personal credit risk of $p2p$ online loans[2] [3], Li Kun and Tian Jiawu et al. Comprehensively compared seven methods of $Lasso-\log istic$ method, $\log istic$ regression, linear discriminant analysis method, k nearest neighbor classification algorithm, support vector machine, BP neural network The algorithm has higher prediction accuracy of the nonlinear classification model[4], while the accuracy and generalization ability of the support vector machine model are slightly better than the logistic regression method [5].

Statistical methods have higher requirements on the original data. In the development of artificial intelligence in recent years, for example, Liu Xiaoya used decision trees and improved DS evidence theory to obtain the final results of credit evaluation[6], which improved the anti-noise ability and generalization ability. Reference[7] uses K-fold cross-validation to optimize the parameters of the support vector machine, but the accuracy is not high. When comparing the genetic algorithm, gray wolf optimization algorithm, and ant colony algorithm to optimize the SVM parameters, the particle swarm algorithm Has the best convergence. In reference [8], the particle swarm optimization algorithm with fast search feature is used to optimize the parameters of the support vector machine, and a better evaluation effect is obtained. In order to solve the shortcomings of the traditional single method in credit evaluation, the combination of models to avoid weaknesses can further improve the accuracy of credit evaluation.

In summary, combining the classification advantages of support vector machines, this paper combines the $C4.5$ decision tree algorithm to perform index screening, and combines the particle swarm optimization algorithm with support vector machines to optimize its parameters.

ISSN: 2688-8653

2. Introduction to Theory

2.1. Support Vector Machine Theory

Support vector machine (SVM) is a supervised learning model and a data analysis model used for classification problems and regression analysis.

For the binary classification problem, it is assumed that there are sample points:

$$X=\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\},y\in\{-1,1\}$$

Among them, the point that determines the maximum distance that the hyperplane can move is called the support vector. The interval is defined as: the sum of the distance between two heterogeneous support vectors to the hyperplane, namely:

$$M=\frac{2}{\|w\|}$$

For the sample points, when the distance between the sample points and the hyperplane is the largest, the expected classification error is the smallest and the classification effect is the best. Under the condition of the largest interval, the constraint conditions are obtained:

$$\begin{cases}\min\phi(x)=\frac{1}{2}w^Tw\\ y_i(w\cdot x_i+b)-1\geq 0\end{cases}$$

Where w represents the normal vector with the direction perpendicular to the hyperplane, x represents the sample point vector, and b is a constant. The above convex quadratic programming problem can be solved according to the duality theory, and the decision function is:

$$f(x)=sign\left(\sum_{i=1}^{l}a_iy_ix_i\cdot x+b\right)$$

x is the sample vector to be tested, x_i is the support vector, and $x_i\cdot x$ is the inner product of the two vectors.

For non-linear problems, the original space can not be better classified with a linear classifier. The original sample is mapped to the p dimension from the original dimension through the kernel function [9] $k(a,b)$. The original sample vector can be separated by the hyperplane in the p-dimensional space. Kernel functions include Gaussian kernel functions, polynomial kernel functions, linear kernel functions, etc. $k(a,b)$ can simplify the inner product operation and satisfy:

$$k(a,b):(a\cdot b+1)^2=\phi(a)\cdot\phi(b)$$

$\phi(a)\cdot\phi(b)$ means that the original sample is mapped into the p-dimensional space for inner product, and $a,b\neq 0$ means any vector. Therefore, the relaxation variable $\xi_i,(\xi_i\geq 0)$ and the penalty parameter C are introduced, and the constraints become:

$$\begin{cases}\min\phi(x)=\frac{1}{2}w^Tw+\left(\sum_{i=1}^{n}\xi_i\right)\\ y_i(w\cdot x_i+b)-1\geq 0\end{cases}$$

According to the dual theory, the decision function is:

$$f(x)=sign(\sum_{i=1}^{l}a_iy_ik(x_i,y_i)+b)$$

This article uses Gaussian kernel functions:

ISSN: 2688-8653

$$k(x,y)=\exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$$

2.2. Theoretical Basis of Particle Swarm Optimization

The particle swarm algorithm (PSO) simulates the bird's foraging behavior. The bird swarm finds the optimal foraging destination through internal cooperation, that is, iteratively searches for the global optimal solution

1. Particle swarm optimization process

Let n particles form a group in the M-dimensional search space, and the position vector of the i-th particle is expressed as:

$$x_i=(x_{i1},x_{i2},...,x_{iM}),i=1,2...n$$

The velocity vector is expressed as:

$$v_i=(v_{i1},v_{i2},...,v_{iM}),i=1,2...n$$

Then the iteration formula of velocity vector is as follows:

$$v_{im}^{k+1}=wv_{im}^{k}+c_1r_1\left(p_{im}^{k}-x_{im}^{k}\right)+c_2r_2\left(p_{gm}^{k}-x_{im}^{k}\right)$$

The iteration formula of the position vector is as follows:

$$x_{im}^{k+1}=x_{im}^{k}+v_{im}^{k+1}$$

Where $i=1,2,...,n;m=1,2,...,M$ is the number of iterations, k is the learning factor, c_1,c_2 is a non-negative constant, r_1,r_2 is a random number in the interval [0,1], w is the determined inertial weight, $v_{im}\in[-v_{\max},v_{\max}]$, $v_{\max}$ is a constant, set according to the actual problem space, p_{im}^{k} and p_{gm}^{k} are the optimal particle position and the global optimal position retrieved by the i th particle iteration number for k times, respectively.

2.3. Theoretical Basis of the Information Gain Method

The information gain method is a method that uses its information entropy for feature selection. By calculating the information gain rate ranking of each indicator, the indicators are further screened.

The calculation formula of the information gain rate is:

$$GainRatio(D,A)=\frac{Gain(D,A)}{SplitInformation(S,A)}$$

Among them, $SplitInformation(S,A)$ is defined as the amount of split information of attribute A.

3. Data Preprocessing

In this paper, the experimental data is selected from 1000 customer data of the German credit database in the public data set of the machine learning library. Through $matlab$ programming, the information gain rate of 20 indicators is calculated as:

ISSN: 2688-8653

Table 1: Information gain rate of each index

index	Information gain rate	index	Information gain rate
A1	22.9	A11	0.01
A2	0.01	A12	0.03
A3	0.01	A13	0.01
A4	4.67	A14	0.04
A5	0.01	A15	2.19
A6	0.06	A16	0.01
A7	20.91	A17	0.02
A8	0.00	A18	0.01
A9	0.01	A19	0.02
A10	0.02	A20	0.01

According to the results, the A_8 index information gain rate is 0, which is not considered and is deleted from the experimental data. Then divide the experimental data data indicators of this article into personal indicators and credit indicators. The data is divided into positive indicators and negative indicators, that is, the larger the positive indicator, the better the impact on the results, and the smaller the negative indicator, the smaller the impact on the results. That is, personal credit indicators: working years (+), gender and marital status (+), current residence years (+), age (+), housing status (+), occupation type (+), telephone (+) credit Indicators: property status (-), other installments (+), other debtors / guarantors (+), savings accounts, debt disposable income percentage (-), credit amount, loan use, current demand deposit status, loan term Credit history (-), the number of current loans in the bank (-), and the number of people to be undertaken (-).

Then use the following data to classify the indicators:

Positive index processing formula:

$$y''_{ij} = \frac{\max\left\{y_{ij}\right\} - y_{ij}}{\max\limits_{1\le i\le n}\left\{y_{ij}\right\} - \min\limits_{1\le i\le n}\left\{y_{j}\right\}} + 1, \left(j = 1, 2...m\right)$$

Negative indicator processing formula:

$$y''_{ij} = \frac{y_{ij} - \min\left\{y_{ij}\right\}}{\max\limits_{1\le i\le n}\left\{y_{ij}\right\} - \min\limits_{1\le i\le n}\left\{y_{j}\right\}} + 1, \left(j = 1, 2...m\right)$$

4. Solving the Model

4.1. Evaluation Index of the Model

According to the research in [1], the classification model evaluation method F score $F-score$ and accuracy rate ($accuracy$) are used to evaluate the model. Among them, $F-score$ is comprehensively considered according to the recall rate and the accuracy rate . The calculation formula is:

$$F-score = \left(1+\beta^2\right)\frac{precision \cdot recall}{\beta^2 \cdot precision + recall}$$

β represents the weight of the recall rate and the accuracy rate, β is greater than one, indicating that the recall rate is more important, otherwise, less than one accuracy rate is more important, this article takes $\beta = 1$.

Among them, the structure matrix of credit classification problem introducing confusion matrix:

ISSN: 2688-8653

$$\begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix}$$

TP means credit is really good, FP means credit is good, FN means credit is bad, and TN means credit is really bad.

Then the recall rate and accuracy rate can be expressed as:

$$recall = \frac{TP}{TP + FN}, precision = \frac{TP}{TP + FP}$$

At the same time, the accuracy rate is expressed as:

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

4.2. Preliminary Study of Optimization Parameters

Set the training set and test set in a 4: 1 ratio, and use $matlab2016a$ to program. The sample data is divided into 800 training sets and 200 test sets. First, according to the support vector machine classifier of the default kernel function parameters, the results can be obtained as:

$$accuracy = 74.5\%, F - score = 46\%$$

4.3. Parameters Optimized by Particle Swarm Optimization

Set the penalty parameter SVM and Gaussian kernel function parameter C of $gamma$ to the initial position x and velocity v. Through the iterative update through the above function, searching for the optimal particle on the basis of determining the constraint conditions, the optimal parameter combination can be screened out.

Use PSO to optimize SVM parameter combination steps:

$step1$: Set the initial value of the particle, that is, the initial value of the penalty parameter C and the parameter $gamma$ of the kernel function, and set the population size of the particle.

$step2$: Evaluate the fitness of each particle and find the initial fitness value. Perform cross-validation first, and use the value of cross-validation as the fitness value.

$step3$: Compare the optimal value of each particle with the fitness value. If it is greater than the fitness value, select the value to replace the fitness value.

$step4$: Compare the fitness value of each particle with the optimal value of the best globally experienced position. If the value is larger, reset the global optimal fitness value and update it according to the above formula.

$step5$: If the condition is met, the optimal $(C, gamma)$ parameter combination is obtained. If the condition is not met, return to x to continue the iteration.

$step6$: Get the optimal $(C, gamma)$ parameter combination, and continue to solve the classification function according to the problem-solving step of svm.

4.4. Determination of Classification Results

The results of each initial value are shown in Table 2:

ISSN: 2688-8653

Table 2: Initial values of particle swarm optimization

parameter	Value	parameter	Value
c_1	1.5	v	3
c_2	1.7	$\max gen$	200
k	0.6	$popc \min$	0.1
w_p	1	$popg \min$	0.001
w_v	1	$pc \max$	100

According to multiple iterations, the result is:

Training set:

$$accuracy = 100\%, F - score = 100\%$$

Test set:

$$accuracy = 96\%, F - score = 87\%$$

Obviously, both the accuracy and the F score are significantly improved. [10]

5. Evaluation and Promotion of the Model

Personal credit evaluation is of great significance to personal credit. Support vector machine is a relatively mature machine learning algorithm. The information gain method and algorithm are used to optimize parameters, which improves the accuracy of the experiment, improves the classification effect of the classifier, and also has better use value. However, this paper only uses algorithms to solve the two-class classification problem, and the multi-class classification problem remains to be studied.

Acknowledgements

This paper was financially supported by foundation: Key Laboratory of Bridge Nondestructive Testing and Engineering Calculation(2019QYY03).

References

[1] Liu Xiaoya, Wang Yingming. Personal credit evaluation model based on C4.5 algorithm to optimize SVM [J]. Computer system application.

[2] Li Jiarong, Jiang Yanli, Tang Liyuan. Personal credit risk assessment of P2P online loans based on BP neural network [J]. Times Finance, 2019 (24): 105-106.

[3] Yang Qiaoyan. Empirical research on credit risk assessment of P2P borrowers based on support vector machine [D]. Shanghai International Studies University, 2018.

[4] Li Kun. Comparative analysis of personal credit evaluation models [J]. Jiangsu Science and Technology Information, 2018, 35 (32): 40-43.

[5] Tian Jiawu. Application of Support Vector Machine and Logistic Regression Model in Personal Credit Prediction [J]. Regional Finance Research, 2018 (11): 25-30.

[6] Liu Xiaoya, Wang Yingming. Research on personal credit evaluation based on support vector machine integration [J / OL]. Computer Engineering and Applications: 1-9 [2020-04-03]. .cn: 999 / kcms / detail / 11.2127.tp.20190716.1522.010.html.

[7] Shangdong Dong. Research on personal credit evaluation based on support vector machine [D]. Beijing University of Chemical Technology, 2017.

[8] Xie Jia. Research on PSO-SVM-based Internet Finance Personal Credit Risk Evaluation Model [D]. Chengdu University of Technology, 2019.

ISSN: 2688-8653

[9] Sun Qiang, Yang Xuna. Application of combined kernel function support vector machine in personal credit evaluation [J]. Heilongjiang Science and Technology Information, 2013 (26): 148-149.

[10] Wei Guannan. Personal credit evaluation based on support vector machine and decision tree CART [J]. Journal of Chengdu Polytechnic University, 2016, 19 (04): 60-62 + 71.