An improved support vector machine and its application in P2P lending personal credit scoring

An improved support vector machine and its application in P2P lending personal credit scoring

Tao Wang [1, a,*] and Jingcong Li [2, b]

[1,2] Information Engineering College, Minzu University of China, Beijing,100081,China

*Corresponding author e-mail:S161199@muc.edu.cn, [b]jcli@pku.org.cn

Abstract. With the help of Internet technologies, P2P(Peer-to-Peer) lending industry has witnessed the rapid development of loan market. From the reason presented above, credit assessment becomes more and more important to the healthy development of P2P load marked. In order to improve accurate predictions of credit assessment, there is necessary to a kind of credit risk evaluation model based on SVM(Support Vector Machines). The performance of SVM depends, to a great extent, on parameters we chose, therefore, our prior work is optimize them. This paper employs an IFOA(Improved Fruit Fly Optimization algorithm) to optimize parameters of SVM model and uses modified model to analyze P2P load data. In the article, we analyze data with four different ways (Linear Regression, Classical SVM, FOA-SVM and IFOA-SVM), and results show that the one presented in this paper has better accurate predictions.

1 Introduction

In some recent years, with the rapid development of Internet technology and its application in financial industry, loan business not only happened in bank but also in many new P2P loan enterprises based on Internet technology. With the development of domestic economy and the change of personal consumption concept, Consumer Loans become more and more acceptable, due to above reasons, Chinese consumer loan business experienced unprecedented development. Comparing to traditional bank loan, P2P have many advantages such as simplified approval process, small amount loan without mortgage, short loan period and so on. Artificial intelligence technology has become a research hotspot, and it has already applied to credit assessment, such as neural network[1][2][3] support vector machines[4][5] and so on. SVM is a novel small sample learning method with a solid theoretical basis with many advantages. For example, a few support vectors determine the final result. The data of loan records containing many privacy is hard to gain, SVM model can reach high performance even with small sample[6]. For these reasons, we choose SVM model to analyze P2P loan data. There are many modified SVM models which had used in credit scoring.

Many international and domestic academics presented a lot of ways to improve SVM. For example, Aiguo Lu(2012)[7] presented an improved SVM based on three-variable working set, which approaches to the optimal solution more quickly, and apply to credit scoring. Lu Han(2017)[8] put

forward orthogonal support vector machine and its application in credit scoring, which effectively solve the problem of dimensional disaster. Harris[9](2015) came up with clustered support vector machine to evaluate credit and it had better classification results.
Danenas and Garsva[10](2015) presented selection of Support Vector Machines based classifiers, which had higher classification accuracy but shortage on stability.

This article mainly optimize parameters of SVM model using an improved fruit fly optimization algorithm and its application in P2P lending personal credit scoring. In experiment, we use four different methods to analyze data which is Linear Regression, Classical SVM, FOA-SVM and IFOA-SVM. The results show that IFOA-SVM model has better accurate predictions.

2 Support Vector Machines

Support Vector Machines was firstly published by Vapnik[11] in 1995, which was based on statistics with a complete theoretical foundation and rigorous theoretical system.

For a given training set $D = \{(\boldsymbol{x}_1,\ y_1), (\boldsymbol{x}_2,\ y_2), \dots, (\boldsymbol{x}_n,\ y_n)\}, y_i \in \{-1, +1\}$, the basic principle of classification learning is to separate the different categories of samples with a divisive hyperplane. The question is that which one should we choose in many divisive hyperplanes in Figure 1 below. Obviously, the black and bold one in the middle is the best, because it has better classification result with the best robustness.

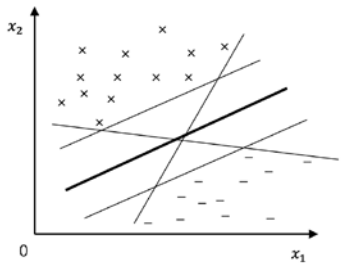

Figure 1 Multiple hyperplanes separate two types of samples

Hyperplanes in the sample space can be described by the following linear equation

$$\boldsymbol{\omega}^T\boldsymbol{x} + b = 0 \tag{1}$$

Where $\boldsymbol{\omega} = (\omega_1, \omega_2, \dots, \omega_n)$ is normal vector, which decides the direction of a divisive hyperplane; b is a constant presented displacement, which is the distance between the divisive hyperplane and origin. Therefore the divisive hyperplane can be expressed as $(\boldsymbol{\omega}, b)$.

The distance between any point in the sample space and the divisive hyperplane can be expressed as

$$d = \frac{|\boldsymbol{\omega}^T\boldsymbol{x} + b|}{\|\boldsymbol{\omega}\|} \tag{2}$$

If the following conditions are met, then hyperplane$(\boldsymbol{\omega}, b)$can separate training samples correctly.

$$\begin{cases} \boldsymbol{\omega}^T\boldsymbol{x}_i + b \geq +1, y_i = +1\ ; \\ \boldsymbol{\omega}^T\boldsymbol{x}_i + b \leq -1, y_i = -1\ . \end{cases} \quad where\ \ (\boldsymbol{x}_i, y_i) \in D \tag{3}$$

The samples which make the equality established are called support vectors. Randomly take two different types of support vectors $(\boldsymbol{\omega}^T\boldsymbol{x}_{i1} + b = +1\ and\ \boldsymbol{\omega}^T\boldsymbol{x}_{i2} + b = -1)$, the sum of distance between them and hyperplane is r, which is called margin.

$$r = \frac{2}{\|\boldsymbol{\omega}\|} \tag{4}$$

In order to find out the optimal hyperplane, we should select the one that makes the margin maximal. Namely, finding out constrained parameters $\boldsymbol{\omega}\ and\ b$, which can meet equation set (3). It can be written in follow standard form:

$$\max_{\boldsymbol{\omega},b} \frac{2}{\|\boldsymbol{\omega}\|} \qquad s.t.\ y_i(\boldsymbol{\omega}^T\boldsymbol{x}_i + b) \geq 1,\ \ i = 1,2,\dots,m \tag{5}$$

Making (5) maximal is equivalent to making $\|\boldsymbol{\omega}\|^2$ minimal. So (5) can be translate to

$$\min_{\boldsymbol{\omega},b} \frac{1}{2}\|\boldsymbol{\omega}\|^2 \qquad s.t.\ y_i(\boldsymbol{\omega}^T\boldsymbol{x}_i + b) \geq 1, \quad i = 1,2,\dots,m \tag{6}$$

This is the basic form of support vector machines.

Formula (6) is convex quadratic programming question, which can be solved by adding Lagrange multiplier $\alpha_i \geq 0$ to each constraint with Lagrange multiplier method. The Lagrange function of formula (6) is

$$L(\boldsymbol{\omega}, b, \boldsymbol{\alpha}) = \frac{1}{2} \|\boldsymbol{\omega}\|^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i(\boldsymbol{\omega}^T \boldsymbol{x}_i + b)) \quad (7)$$

We can get $\boldsymbol{\omega}$ and b by solving formula (7), and the model corresponding to maximum margin hyperplane is

$$f(\boldsymbol{x}) = \boldsymbol{\omega}^T \boldsymbol{x} + b = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i{}^T \boldsymbol{x} + b \quad (8)$$

There is another problem that you might find a few samples of positive class in negative class set and vice versa. In order to solve this problem, there is necessary to introduce a concept of soft margin, namely, it allows several samples not meet the constraint condition (3). And for linearly inseparable cases, We can adapt formula (6) for (10) with adding slack variable $\xi_i \geq 0$ and penalty factor C in constraint condition.

$$\min_{\boldsymbol{\omega}, b, \xi_i} \frac{1}{2} \|\boldsymbol{\omega}\|^2 + C \sum_{i=1}^{m} \xi_i \qquad s.t.\ y_i(\boldsymbol{\omega}^T \boldsymbol{x}_i + b) \geq 1 - \xi_i,\ \ \xi_i \geq 0, i = 1,2,\dots,m \quad (9)$$

C stands for penalty of misclassification, larger C means better fitting effect; ξ_i for the optimization of the generalization capacity.

In nonlinearity cases, linearly inseparable questions need to translate to linearly separable with kernel methods which are concerned with mapping original space into a higher dimensional vector space. SVM model then will be rewrite into

$$f(\boldsymbol{x}) = \boldsymbol{\omega}^T \phi(\boldsymbol{x}) + b = \sum_{i=1}^{m} \alpha_i y_i \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}) + b = \sum_{i=1}^{m} \alpha_i y_i k(\boldsymbol{x}, \boldsymbol{x}_i) + b \quad (10)$$

Obviously, penalty factor C and kernel function k are the most important parameters which have significant impact on performance of SVM. This paper mainly adopt improved fruit fly optimization algorithm to optimize them.

3 Improved Fruit Fly Optimization Algorithm

Fruit Fly Optimization Algorithm [12][13] (short for FOA) was firstly presented by Wenchao Pan, a Taiwan scholar, and used to optimize generalized regression neural network and its application in evaluating the operation performance of company.

Fruit Fly Optimization Algorithm is a new way to find global optimal solution based on foraging behavior of fruit fly. Fruit fly is superior to other species in sensory perception, especially in smell and vision. And its olfactory organs are good at collecting all kinds of odors in the air, its visual organs finding the position of foods and its group. Then they will fly to that position.

Although fruit fly optimization algorithm already has successfully apply in many ways, there are still some defects needed to be solved[14], such as slow convergence rate, lower algorithm accuracy, pre-mature convergence phenomenon and so on.

Comparing to traditional FOA which only seek one optimal fruit fly position, the IFOA theory is to find multiple optimal fruit fly positions, then weight these positions to get an optimized weighted position.

The steps of IFOA can be described as follows:

Step01: Initialize(X_0, Y_0), the position of fruit fly group; m, the number of fruit fly; and n the

number of iterations.

$$X(i,:) = X_0 + RandomValue; \qquad Y(i,:) = Y_0 + RandomValue. \qquad (i = 1,2,\dots,m)$$

Step02: Calculate the distance between original point and fruit fly, $Dist(i,:)$ means the distance between ith fruit fly and original point. Then get the smell concentration judgment value $S(i,:)$ according to reciprocal of Dist(i,:).

$$Dist(i,:) = \sqrt{[X(i,:)]^2 + [X(i,:)]^2} \quad , \quad S(i,:) = \frac{1}{Dist(i,:)}$$

Assigning values to penalty factor C and kernel function k

$$C = 10S(i); \qquad K = S(i)$$

Step03: Substitute in $S(i,:)$ fitness function to get the smell concentration $Smell(i)$ of the position of the ith fruit fly.

$$Smell(i) = F(i) = \frac{1}{3}\sum_{i=1}^{3}\sqrt{\frac{1}{n}\sum_{j=1}^{n}\left[f(x_{ij}) - y_{ij}\right]^2}$$

Where y_{ij} stands for the actual value, $f(x_{ij})$ the predicted value of cross verification, n the number of fruit fly of each sub-training sample in cross verification.

Step04: Find out n elite fruit flies which have highest smell concentration, and weight their positions to get an optimized weighted position $\bar{b}$.

$$\left[b_{Smell(i)}, b_i\right] = \max\left(Smell(\mathrm{i})\right), i = 1,2,\dots,n; \qquad \bar{b} = (b_1 + b_2 + \cdots + b_n)/n$$

Step05: Fruit fly group fly to position $\bar{b}$ according to collected visual information, this step is performed n times and find out the best one to next step.

Step06: Save the best smell concentration Smell and its coordinates, the rest of fruit fly group will fly to the position to create a new cluster location according to visual information.

$$Smell_{best} = b_{Smell}; \qquad X_0 = X(\bar{b}); \qquad Y_0 = Y(\bar{b})$$

Step07: Enter iterative process, repeat step03 to step05, estimate if current smell concentration is superior to last iteration, if so, execute step06; if not end the loop when performed n times iteration, and output the optimal value.

4 Experimental design and results

The data used in this paper are from the website of RenRenDai, a load platform set up in 2010, and they are true loan information and hard to obtain. Therefore, we only have 1000 samples, the half as training samples and the other half as test samples. In training samples, there are 362 risk-free samples and 138 risk samples. In test samples, there are 315 risk-free samples and 185 risk samples.

Firstly, we need to establish IFOA-SVM model used LIBSVM software package, the flow chart is described in figure 2 below. The core problem in this article is to optimize parameters of SVM used IFOA. The steps of how to optimize parameters of SVM are as follows, and its flow chart is in figure3.

Step01: Select moderate function (Mean Square Error).

Step02: Initialize the numbers of fruit fly group(20) and iterations(50); parameters of SVM.

Step03: Establish SVM model and forecast result, record the best value of each iteration.

Step04: Update the newest position of fruit fly based on result of above steps. Repeat step01 to step03 until reaching to the maximum number of iterations, output optimal parameters C and g.

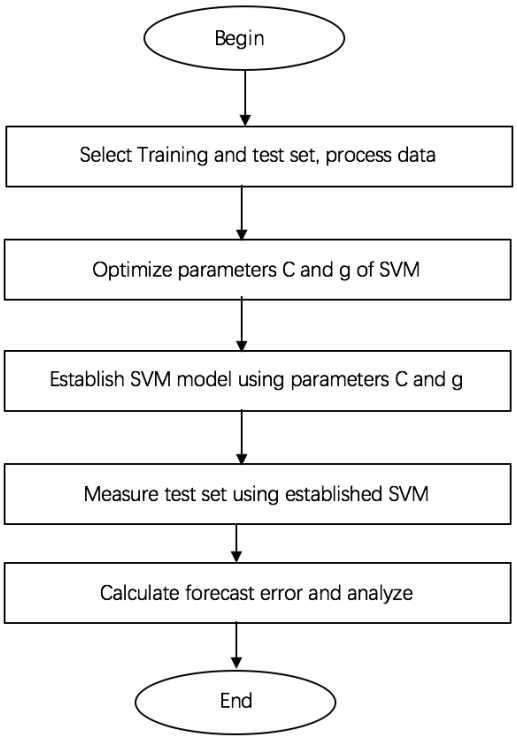

Figure 2 Flow chart of establishing IFOA-SVM

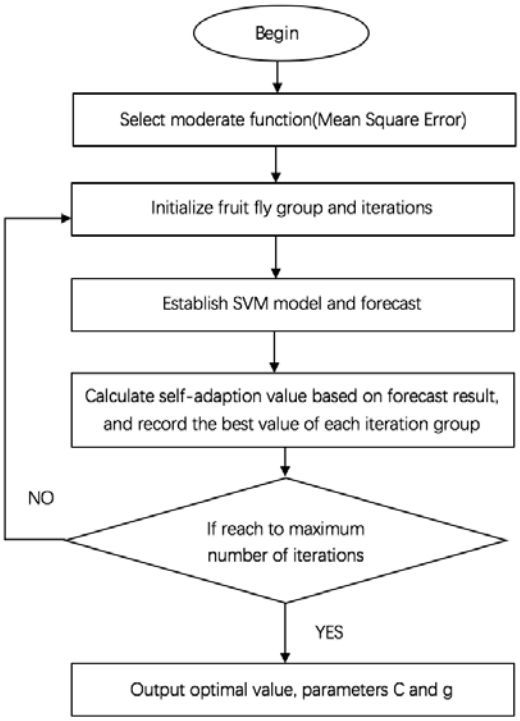

Figure 3 Optimize SVM parameters using IFOA

The following are the results of 4 sets of experiments on the same data, in order to see the results more clearly, we only intercept the first 100 comparison of actual results and forecast results. The symbol 'o' stands for the actual results of test samples, the symbol '△' stand for predicted results. In the results of four group experiences, there are three different type situations. The first one is that two symbols overlap together, which is the correct one. The second is only symbol 'o' in the ordinate equal to 1, which means that the kind of model mistakenly divide the positive one into negative one. The third is opposite to the second, it mistakenly divides the negative one into positive one.

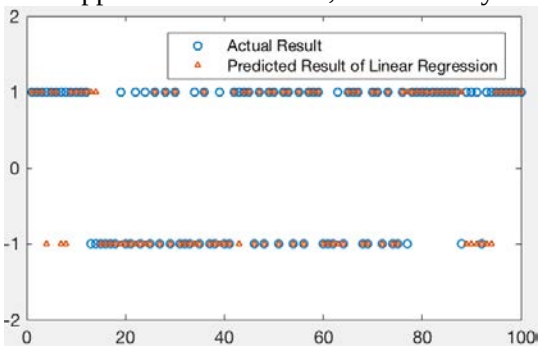

Figure 4 Linear Regression Model

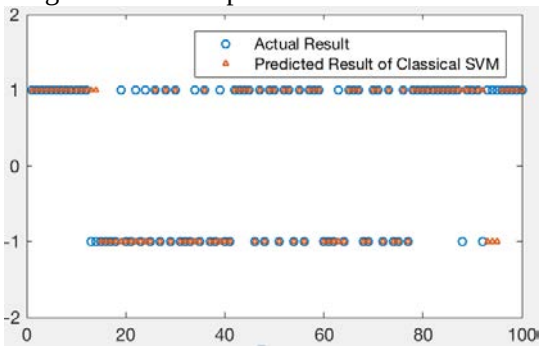

Figure 5 Classical SVM Model

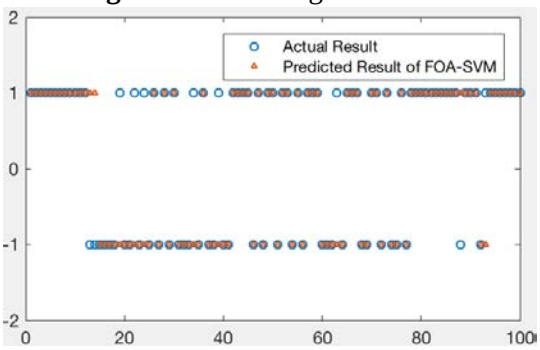

Figure 6 FOA-SVM Model

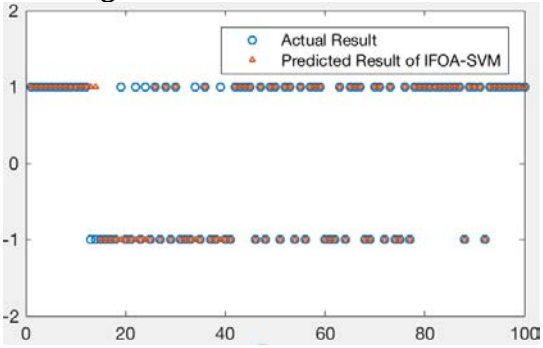

Figure 7 IFOA-SVM Model

The classification Results of four Prediction Models are list in Table 1 below. From the result, we can draw a conclusion that IFOA-SVM model has higher precision rate than other three groups, and low the number of misjudgment and rise precision rate used IFOA optimizing SVM.

Table 1 Classification Results of four Prediction Models

	Test Sample	Error	Precision Rate
Linear Regression Model	500	112	77.6%
Classical SVM Model	500	88	82.4%
FOA-SVM Model	500	46	90.8%
IFOA-SVM Model	500	35	93%

5 Conclusion

With the development of lending market, the research of credit risks will get more and more attention. In this paper, there are four group experiences used to analyze P2P lending data, and experimental results show that IFOA-SVM model has better classification result than other ways. There is also some room to be improved in this article, for example, if we can combine the data with our existing data, which consumers purchase things with credit products such as Ant Check Later to Alipay, Baitiao to Jingdong and so on, this will make our data more multiple and enable us to get more accurate predictions.

References

[1] Zhao Z, Xu S, Kang B H, et al. Investigation and improvement of multi-layer perceptron neural networks for credit scoring[J]. Expert Systems with Applications, 2015, 42(7):3508-3516.
[2] Tavana M, Abtahi A R, Caprio D D, et al. An Artificial Neural Network and Bayesian Network Model for Liquidity Risk Assessment in Banking[J].Neurocomputing,2017,In Press:2525-2554.
[3] Khashman A. Credit risk evaluation using neural networks: Emotional versus conventional models[J]. Applied Soft Computing, 2011, 11(8):5477-5484.
[4] Dikkers H, Rothkrantz L. Support vector machines in ordinal classification: An application to corporate credit scoring[J]. Neural Network World, 2005, 15(6):491-507.
[5] Huang C L, Chen M C, Wang C J. Credit scoring with a data mining approach based on support vector machines[J]. Expert Systems With Applications, 2007, 33(4):847-856.
[6] Panja R, Pal N R. MS-SVM: Minimally Spanned Support Vector Machine[J]. Applied Soft Computing, 2018, 64:356-365.
[7] Aiguo Lu, Jue Wang, Hongwei Liu. An improved SVM learning algorithm and its applications to credit scoring [J].Systems Engineering – Theory & Practice, 2012, 32(03): 515-521.
[8] Lu Han, Liyan Han. Orthogonal support vector machine and its application in credit scoring[J].Journal of Industrial Engineering Managemeng,2017,31(02):128-136.
[9] Harris T. Credit scoring using the clustered support vector machine[J]. Expert Systems with Applications, 2015, 42(2):741-750.
[10] Danenas P, Garsva G. Selection of Support Vector Machines based classifiers for credit risk domain[J]. Expert Systems with Applications, 2015, 42(6):3194-3204.
[11] Vapnik V N. The Nature of Statistical Learning Theory [M]. Berlin: Springer 1995.
[12] Pan，W T.A new evolutionary computation approach: Fruit Fly Optimization Algorithm[C]. 2011Conference of Digital Technology and innovation Management Taipei，2011.
[13] Pan，W T.A new fruit fly optimization algorithm: Taking the financial distress model as an example[J]. Knowledge-Based Systems，In Pres，2011.
[14] Gangquan Si, Shuiwang Li, Jianquan Shi, Zhang Guo. Least Squares Support Vector Machine Parameters Optimization Based on Improved Fruit Fly Optimization Algorithm with Applications[J].Journal of Xi'an Jiaotong University,2017,51(06):14-19.