IOP Conf. Series: Materials Science and Engineering **439** (2018) 042003 doi:10.1088/1757-899X/439/4/042003

# Personal credit assessment based on improved SVDD algorithm

**Ting-ting Dai**

(Zhaotong College, School of Mathematics and Statistics, Yunnan, Zhaotong 657000)

**Abstract:** Aiming at the credit evaluation problem of commercial banks, this paper proposes a new personal credit evaluation method based on the existing SVDD algorithm--the improved SVDD algorithm, and introduces the theory and derivation of the algorithm in detail, based on this, an example is given.and tested, the results show that the improved SVDD effect is indeed better than SVDD.

## 1. Introduction

At present, mainstream credit research mainly considers personal credit evaluation as three issues: clustering problem, classification problem, and regression problem. Based on the above three questions, the current research methods of personal credit assessment mainly includes neural networks,[1] The method is to model the nonlinear prediction model of the biological neural network structure and learn the pattern recognition.Genetic algorithm[2] The structure of biological neural network is a natural selection mechanism that borrows from the natural world, "survival of the fittest, survival of the fittest" and the calculation method of random search optimal solution evolved from biological genetic characteristics.Support Vector Machine [3] (SVM), both the method and the neural network belong to the artificial intelligence domain calculation method. In addition to the same functions as the neural network, it can also make up for the shortcomings of the neural network method. In this paper, an improved SVDD algorithm is obtained based on the standard SVM and SVDD, and the personal credit is evaluated by this method.

## 2. Support vector data description

Support Vector Data Description (SVDD) was proposed by Tax et al[4-5] . in 1999 on the basis of the theory of minimum bounding sphere (MEB) and support vector machine (SVM). The basic idea is: Find a super-autumn sphere that encloses all or most of the target sample in the feature space, and minimize the volume enclosed by the hypersphere so that the target sample is surrounded by the hypersphere as much as possible, rather than the target sample. As far as possible, it is not included in the hypersphere, thus achieving the division of categories.

### *2.1. Improved support vector data description derivation*

Support vector data description is a special SVM, but it is different from SVM. It is a kind of supervised learning method. When doing credit evaluation modeling, sometimes it is not known whether the customer is a "good credit" customer or a "credit difference". The client, at this time unsupervised learning methods, then unsupervised learning methods are effective means of supervision. However, under normal circumstances, unsupervised learning accuracy will be lower.

Therefore, this paper presents an improved SVDD method based on SVDD to improve the accuracy of its learning.

The objective function in the original problem of standard SVDD is:

$$R^2 + C\sum_{i=1}^{n}\xi_i \tag{1}$$

Therefore, this paper presents an improved SVDD method based on SVDD to improve the accuracy of its learning.

The objective function in the original problem of standard SVDD is:

$$R^2 + C\sum_{i=1}^{n}\xi_i^2 \tag{2}$$

$$s.t.\left\|\varphi(x_i)-a\right\|^2 \le R^2+\xi_i, i=1,2,...,n \tag{3}$$

among them, $\xi_i$ as a slack variable, $a$ is a column vector,Indicates the center of the ball that is being sought. $\varphi(\cdot)$ is a feature function that maps a sample from the input space to the feature space the real. $C>0$ is a penalty parameter, The bigger the $C$, the greater the penalty for misclassification.By setting the parameter $C$,The compromise between the radius R of the hypersphere and the number of training samples it can contain is made. When the $C$ value is large, try to put the sample into the goal; when the $C$ value is small, try to compress the size of the ball. It is worth noting that the constraint $\xi_i \ge 0$ is not needed in the improvement of SVDD, because it is assumed that there is $\xi_i < 0$ for some of the za8 target values, and $\xi_i^* = 0$ is satisfied, satisfying the condition:

$$\left\|\varphi(x_i)-a\right\|^2 \le R^2+\xi_i^* \le R^2+0 \tag{4}$$

Make the target value smaller, which contradicts $\xi_i$ as the optimal solution.

Introducing the Lagrange multiplier $\alpha_i$, the Lagrange function corresponding to the problem (2)(3) is:

$$L = R^2 + \sum_{i=1}^{n}\xi_i^2 - \sum_{i=1}^{n}\alpha_i(R^2+\xi_i-\left\|\varphi(x_i)-a\right\|^2) \tag{5}$$

In equation (5), L is deduced relative to *R, a,* $\xi_i$ and is equal to zero.

$$\frac{\partial L}{\partial R} = 2R - 2R\sum_{i=1}^{n}a_i = 0 \Rightarrow \sum_{i=1}^{n}a_i = 1 \tag{6}$$

$$\frac{\partial L}{\partial a} = -2\sum_{i=1}^{n}\alpha_i(\varphi(x_i)-a) = 0 \Rightarrow a = \sum_{i=1}^{n}\alpha_i(\varphi(x_i) \tag{7}$$

$$\frac{\partial L}{\partial \xi_i} = 2C\xi_i - \alpha_i = 0 \Rightarrow \alpha_i = 2C\xi_i \tag{8}$$

Bringing the equations (6)-(8) into the Lagrange function (5), the dual problem of the optimization problem is obtained:

$$\max \quad L = \sum_{i,j=1}^{n}\alpha_i K(x_i,x_j) - \sum_{i,j=1}^{n}\alpha_i\alpha_j(K(x_i,x_j)+\frac{1}{4C}\delta_{ij}) \tag{9}$$

$$s.t.\sum_{i=1}^{n}\alpha_i = 1, i=1,2,...,n \qquad 0 \le \alpha_i, i=1,2,...,n$$

$$\text{among them, } \delta_{ij} = \begin{cases} 1, i = j \\ 0, i \neq j \end{cases} \tag{10}$$

It can be seen from the derivation process and the dual problem that the C value does not directly affect the constraint of the dual problem as the standard SVDD, but also directly affects the quadratic term of its objective function. In addition, the improved SVDD is a form of the standard SVDD at the quadratic loss function, which is constructed by reducing the risk from the primary of the deviation to the quadratic.

According to the optimization theory, the KKT condition for the improved SVDD problem is:

$$\alpha_i(R^2 + \xi_i - \|\varphi(x_i) - a\|^2) = 0, \quad \|\varphi(x_i) - a\|^2 \leq R^2 + \xi_i, \quad \alpha_i = 2C\xi_i \quad i = 1, 2, ..., n \tag{11}$$

According to the difference of $\alpha_i$, these conditions can be further refined, that is, divided into the following cases:

(1): When $\alpha_i = 0$ $\xi_i = 0$ $\|\varphi(x) - a\|^2 \leq R^2 \Leftrightarrow \alpha_i = 0$, there is , therefore .

(2): When $\alpha_i > 0$ $\xi_i > 0$ $\|\varphi(x) - a\|^2 > R^2 \Leftrightarrow \alpha_i > 0$, there is , therefore .

By the equation:

$$a = \sum_{i=1}^{n} \alpha_i \varphi(x_i) \tag{12}$$

The center of the ball is given as a linear combination of the data $x_i$, and $\alpha_i$ is the solution to the dual problem.

The decision function is as follows : $f(x) = R^2 - \|\varphi(x) - a\|^2$ among them,

$$\|\varphi(x) - a\|^2 = K(\varphi(x), \varphi(x)) - 2\sum_{i=1}^{n} \alpha_i K(\varphi(x), \varphi(x_i)) + \sum_{i,j=1}^{n} \alpha_i \alpha_j K(\varphi(x_i), \varphi(x_j)) \tag{14}$$

R is the distance from any support vector to the center of the sphere on the hypersphere.

When $f(x) \geq 0$, the corresponding point is judged to be a normal point, otherwise it is judged to be an abnormal point.

## 3. Algorithm

From the above derivation, the improved SVDD algorithm can be derived as follows:

**Step 1:** Enter data set $X = \{x_i\}$, $x_i \in R^1$, $i = 1, 2..., n$, and standardize the data using the Z-scoreg method;

**Step 2:** Select the appropriate kernel function $K(x_i, x_j)$ (eg, Gaussian kernel), and use cross-validation to select parameters C and g;

**Step 3:**: construct and solve the optimization problem (9), and find the optimal solution $a^* = (a_1^*, a_2^*, ..., a_n^*)$;

**Step 4**: Select the sample point $x_l$, $l = 1, 2..., k$ corresponding to the positive component $a_1^*$ of

$a^*$. Calculate $R^2 = \frac{1}{k}\sum_{l=1}^{k}[1-2\sum_{i=1}^{n}\alpha_i^* k(x_i,x_l)+\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j K(x_i x_j)], a=\sum_{i=1}^{n}\alpha_i^* x_i$ .

**Step 5**: constructing the decision function $f(x)=R^2-\|x-a\|^2$. if $f(x)>0$, then the sample point is in the hypersphere; if $f(x)<0$, the sample point is beyond the request;

**Step6**: Classify the sample and output the sample category.

## 4. Data simulation experiment

The improved SVDD algorithm is used to address the corresponding issues in personal credit assessment. We consider the customer sample point of 'good credit' as a normal point, and consider the customer sample point of 'credit difference' as an abnormal point to establish a credit evaluation of the SVDD model of unsupervised learning.

In order to compare performance, the creditworthiness model of SVDD and improved SVDD was established using comparative authoritative German credit data to verify the effectiveness of the improved SVDD. The experiment is mainly carried out by programming with MATALBR2012a. The kernel function $k(x,x_i)$ of all models adopts Gaussian radial basis kernel function.

$$K(x,x_i)=\exp(-\frac{\|x-x_i\|^2}{2g^2}) \tag{15}$$

### *4.1 Determination of parameters C and g*

In the test, all the parameters are 5-fold cross-validation [6] grid optimization method, the penalty parameter C and the kernel parameter g are in the search range $\{2^{-7},...,2^4\}$ and $\{2^{-3},...,2^3\}$, respectively, the initial value C=1, g=0.1, 5 - The folded cross-validation grid optimization method selects the optimal (C, g) pair value, and the optimal parameter in the German credit data [7] evaluation model is C=10.5561, g=6.6029 (as shown in Figure 1). In order to better reflect the relationship between (C, g) pairs and accuracy, Figure 2 gives its corresponding 3D view. Once the parameters are determined, the data can be substituted for credit evaluation analysis.

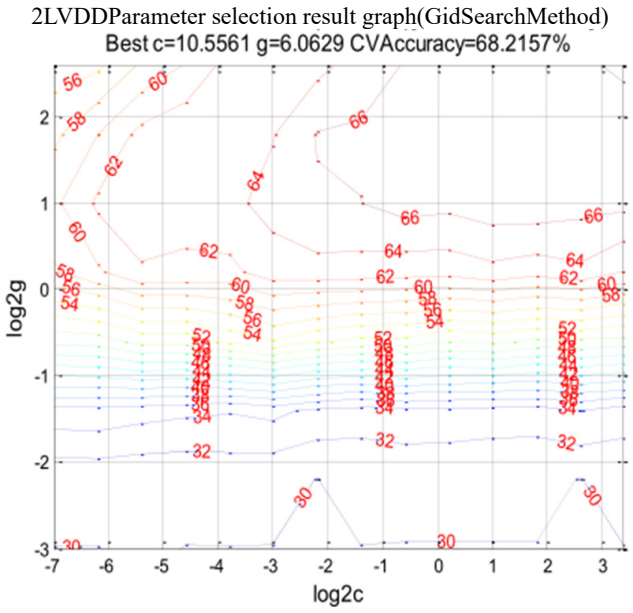


Figure 1 Germany credit data improved SVDD model parameter selection results contour map

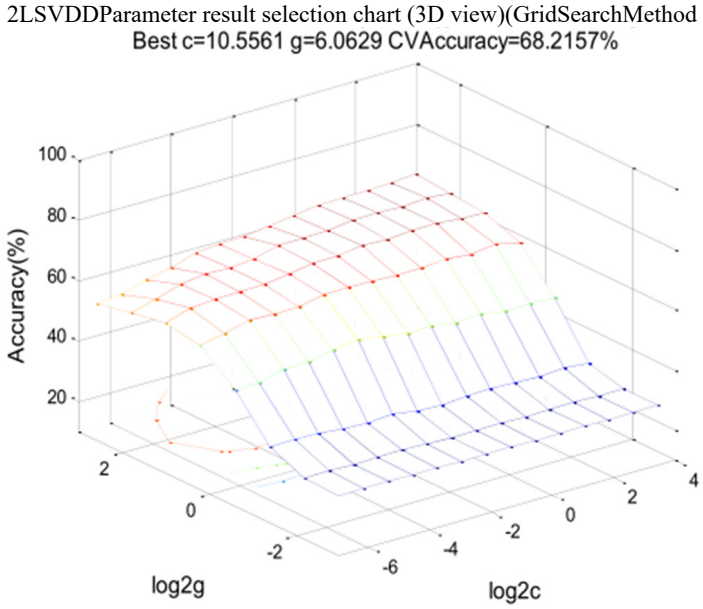


Figure 2 German credit data improved SVDD model parameter selection results 3D view

*4.2 Test results and analysis*

After the parameters are determined, the test is conducted. The test results are shown in Table 1:

Table 1 German credit data test results

| algorithm | Accuracy(%) | |
|---|---|---|
| | Training set | Test set |
| SVDD | 49.18 | 42.64 |
| Improved SVDD | 66.72 | 68.77 |
| SVM | 81.4093 | 76.5766 |

Table 1 shows the test results of German credit data. The test results of the improved SVDD model are: the accuracy of the training set is 66.72% and the test set is 68.77%, which is obviously better than the SVDD test result, but lower than the SVM test result.

Experimental results show that the improved SVDD effect is indeed better than SVDD, but the improved SVDD, like SVDD, is an unsupervised learning method with lower performance than the standard SVM.

## 5. Conclusion

Based on the standard SVDD, this paper proposes a new algorithm: an improved SVDD algorithm, using credit evaluation data for data experiments, comparing with the performance of SVDD and standard SVM, and obtaining improved SVDD effect. It is indeed better than SVDD

## References

[1] The ECB stress test revealed amazing data: European bad debts up to 100 million [N/OL].2014.10-27
[2] Zhang Qiao. The bad debt rate of Japanese student loans surged and the debts of debt students were difficult to repay [G/OL].2013.05-08
[3] One Fortune Xiaobian. The bank is behind the dance: the risk of bad debts has not been increased [N/OL].2014.11-8

[4] Tax D M J,Duin R P W.Data Domain Description using Support Vectors[J].In proceeding of the 7th European Symposium on Artificial Neural Networks Bruges(Belgium)Bruges,1999, 251-257.
[5] Tax D M J,Duin R P W. Support Vector domain Description[J].pattern Recognition Letters, 1999,20:1191-1199
[6] YAN D, NG W L,ZHANG X, etal,Targeting DNA-PKcs and ATM with miR-101 sensitizes tumors to radiation[J].plos One,2010,5(7):e11397.
[7] http://archive.ics.uci.edu/ml.