Customer Credit Scoring Method Based on the SVDD Classification Model with Imbalanced Dataset

Bo Tian[1], Lin Nan[2], Qin Zheng[1], and Lei Yang[3]

[1] School of Information Management & Engineering, Shanghai University of Finance and Economics, Shanghai, 200433 China
[2] School of International Business Administration, Shanghai University of Finance and Economics, Shanghai, 200433 China
[3] School of Economics and Management, Tongji University, Shanghai 200092, China
youngtb@gmail.com, yangyoungya@sina.com

Abstract. Customer credit scoring is a typical class of pattern classification problem with imbalanced dataset. A new customer credit scoring method based on the support vector domain description (SVDD) classification model was proposed in this paper. Main techniques of customer credit scoring were reviewed. The SVDD model with imbalanced dataset was analyzed and the predication method of customer credit scoring based on the SVDD model was proposed. Our experimental results confirm that our approach is effective in ranking and classifying customer credit.

Keywords: Customer Credit Scoring, Pattern Classification, Support Vector Domain Description Model, Imbalanced Dataset.

1 Introduction

Pattern classification and predication is one of main problems in statistical decision, pattern recognition and artificial intelligence, signal detection and estimation. Classical statistical methods of pattern classification mainly include Bayes statistical discriminate method, Fisher discriminate method, log-linear regression model and so on [1-3]. Numbers of samples in classical statistical methods are usually assumed to be sufficiently large. But samples usually are finite even deficient in practice. So in recent years, artificial intelligence methods such as neural-networks, clustering method, support vector machine (SVM) model which are based on finite samples become more and more popular in pattern classification field [4-6].

Customer credit scoring of banks is also called default risk prediction, which is predicting the possibility of losing that banks are suffered because the customers of banks are reluctant or unable to fulfill the credit contract [7]. Evaluation and predication of customer credit of banks is to predict the probability that customers repay loans on schedule based on all kinds of information that customers offered, and decide whether or not to approve the loan applications of the customers. Customer credit scoring is a powerful tool for the management of credit risk of banks which is a

M. Zaman et al. (Eds.): CETS 2010, CCIS 113, pp. 46–60, 2010.

kind of pattern classification problem. In early time, the main objects of credit scoring are small and medium enterprises such as shopkeepers, mail corporations, financial firms. And the numeric graded credit scoring decision systems and statistical classification techniques are used. Individual credit card service appeared in 1960s. The numbers of individual customers and the total consumption amounts exceeded the small and medium enterprises gradually. The credit scoring decision-making processes should be performed automatically. And the development of computer and Internet techniques provided technical guarantee for the automatization of decision-making. Following, statistical methods such as Bayes discriminate method, Fisher discriminate method, and log-linear regression model were applied in the processes of individual customer credit scoring of banks [1, 3, 7]. The customers credit scoring methods based on classical statistical theories are usually under the hypothesis of asymptotic theory that the numbers of the samples were prone to infinite. And prior knowledge such as probability distributions of the samples and properties of the estimators about the samples are used. In 1990s, data-driven artificial intelligence classification methods such as neural-networks and SVM models were introduced in the analysis processes of customer credit scoring of banks [7-9]. The data-driven classification methods are based on the statistical learning theory. The disadvantages of statistical asymptotic theory can be tided over using the data-driven classification methods. The principle of the data-driven classification methods is that the decision-making function is achieved by the learning process using small or finite samples about the objects. And prior knowledge about the samples needed not to be known. Minimization of experimental risk is used in neural- networks predication models, which make the total output error be minimized. Individual credit scoring method based on the neural-networks model was investigated in reference [8]. But the practical applications of the neural-networks are limited because of several shortcomings such as over-fitting phenomenon in learning processes, lack of generalization ability, and local extremum values. Cortes and Vapnik proposed the SVM model in which the decision super-plane is constructed by minimizing of structural risk, and the complexity of models and experimental risk are balanced effectively [10-11]. The SVM models have strong generalization ability. And problems such as small number of samples, non-linear map, high dimension description, and local extremum values can also be solved. So the SVM models are very suitable to be used in pattern classification with small samples, approximation of functions and so on. Some improved SVM models were then proposed by other researchers. Least squares SVM model was proposed by Suykens [12]. Wavelet SVM model was proposed by Zhang [13]. Support vector domain description (SVDD) model was proposed by Tax [14]. Applications of the SVM models were also investigated. Individual credit scoring method based on the SVM model was investigated in reference [15]. Different costs of misclassification were considered in the cost sensitive SVM model proposed by Zadrozny [16]. Classification algorithm of SVM model with imbalanced dataset was proposed in [17].

Evaluation and predication of customer credit of banks are becoming more and more important with the development of individual credit card service in commercial banks. The object of the management of credit risk turns to maximizing the profit of

commercial banks from minimizing the probability of breach of contracts [7]. Customer credit scoring of banks is a kind of pattern classification problem with imbalanced dataset because there exists obvious discrimination of sample numbers between two classes named well-record and bad-record [18]. It is subjective to determine the weights of positive and negative samples manually in the weighted SVM model in reference [17]. An improved SVM model, the SVDD model which based on dataset description method was proposed by Tax [14]. The SVDD model initially deals with the problem of one-class classification [19]. The principle of the SVDD model is to construct the hyper-sphere with minimizing the radius which contains the most of positive examples, and others samples named outliers are located outside of the hyper-sphere. The computing tasks of the SVDD model are to calculate the radius and center of the hyper-sphere by using the given samples. And the SVDD model can be used in describing the dataset and detecting outliers. The dataset are described by using the samples located at boundary of hyper-sphere in the SVDD model, which makes the SVDD model to be high computing efficient. As we analyzed, customer credit scoring of banks is a pattern classification problem with imbalanced dataset. New predication method of customer credit scoring of banks based on the SVDD model with the imbalanced dataset was proposed in this paper. And residual of the paper was organized as follows. Main techniques of customer credit scoring were reviewed in section one. The SVDD classification model with imbalanced dataset was analyzed in section two. And the multiplicative updating principle to compute the parameters of the model was also discussed in this section. Then predication method of customer credit scoring of banks based on the SVDD model was proposed in section three. Section four reported the experimental comparing results of artificial dataset and benchmark credit dataset of banks using the proposed method and the SVM- based method. Conclusions were drawn in the last section.

2 The SVDD Model with Imbalanced Dataset

An optimal closed high-dimensional hyper-sphere is established in the SVDD model as for classification problems. Positive examples are included in the hyper-sphere, and outliers are located outside of the hyper-sphere. And the SVDD model can be used in describing dataset and detecting outliers. Following the primary SVDD model with one class of samples is reviewed. And the multiplicative updating principle to compute the parameters of the model is analyzed. Then the SVDD model which containing two classes of samples was discussed, and the decision-making function based on the hyper-sphere of the SVDD model is shown. Kernel transformation in feature space is also mentioned in this section.

2.1 The SVDD Model to Describe One Class of Samples

Let dataset $\{x_i, i = 1, 2, ...N\}$ be the known samples, where N is the number of samples. The SVDD model is used to describe the dataset. The parameters of the

SVDD model are the radius R and the center $\mathbf{a}$ of the hyper-sphere. The object is to describe the dataset by using a hyper-sphere with minimized radius in feature space. In other words, all the samples should be located in the hypersphere. The mathematical form of the model is minimizing the function $F(R,\mathbf{a}) = R^2$ with the constraint condition $\|x_i - \mathbf{a}\|^2 \le R^2 (\forall i = 1,2,...,N)$. If outliers appear in the dataset, the distances from the samples $\{x_i, i = 1,2,...,N\}$ to the center of the hyper-sphere were not strictly smaller than R. But large distance should be penalized. Thinking the influence of outliers, slack variable $\xi_i \ge 0,\ (i = 1,2,...,N)$ are introduced in the objective function. So the problem of minimizing the radius of the hyper-sphere can be described by the following quadratic programming with inequality constraints

$$\begin{cases} \min R^2 + C\sum_{i=1}^{N} \xi_i \\ sub: \|x_i - \mathbf{a}\|^2 \le R^2 + \xi_i, \xi_i \ge 0, i = 1,2,...,N. \end{cases} \tag{1}$$

where the positive constant parameter C is called penalty factor. The parameter C controls the tradeoff between the radius of the hyper-sphere and the error.

Using the Lagrange multiplier algorithm for Eq.(1), the corresponding Lagrange function is

$$\begin{aligned} &L(R,\mathbf{a},\alpha_i,\beta_i,\xi_i) = \\ &\quad R^2 + C\sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i (R^2 + \xi_i - \|x_i - \mathbf{a}\|^2) - \sum_{i=1}^{N} \beta_i \xi_i \end{aligned} \tag{2}$$

where $\alpha_i \ge 0, \beta_i \ge 0$ are Lagrange multipliers. Lagrange function L should be minimized with respect to R, $\mathbf{a}_i$, ξ, and maximized with respect to α_i and β_i. The extremum conditions of Lagrange function L are

$$\frac{\partial L}{\partial R} = 0, \frac{\partial L}{\partial \mathbf{a}} = 0, \frac{\partial L}{\partial \xi_i} = 0 \tag{3}$$

such that

$$\sum_{i=1}^{N} \alpha_i = 1 \tag{4}$$

$$\mathbf{a} = \sum_{i=1}^{N} \alpha_i x_i \tag{5}$$

$$C - \alpha_i - \beta_i = 0 \tag{6}$$

We can get $0 \le \alpha_i \le C$ from Eq.(6) because $\alpha_i \ge 0, \beta_i \ge 0$. When Eq.(4-6) are substituted into Lagrange function Eq.(2), the dual form of the Lagrange optimization problem turns into

$$\begin{cases} \max \sum_{i=1}^{N} \alpha_i (x_i \cdot x_i) - \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j (x_i \cdot x_j) \\ sub: \sum_{i=1}^{N} \alpha_i = 1, \quad 0 \le \alpha_i \le C, \quad i, j = 1, 2, ..., N. \end{cases} \tag{7}$$

where $x_i \cdot x_j$ is the inner product of x_i and x_j. Usually the dataset is not distributed in the hyper-sphere ideally. So the inner product can be substituted by some kernel function in high-dimensional feature space. After solving the quadratic programming problem containing inequality constraints denoted by Eq.(7), the parameters of the SVDD model $\{\alpha_i, i = 1, 2, ...N, \}$ is achieved. The parameters satisfy the following conditions

$$\begin{cases} \left\| x_i - \mathbf{a} \right\|^2 < R^2 \rightarrow \alpha_i = 0 \\ \left\| x_i - \mathbf{a} \right\|^2 = R^2 \rightarrow 0 < \alpha_i < C \\ \left\| x_i - \mathbf{a} \right\|^2 > R^2 \rightarrow \alpha_i = C \end{cases} \tag{8}$$

The multiplicative updating algorithm to solve the quadratic programming problem containing inequality constraints denoted by Eq.(7) of the SVDD model will be analyzed in next subsection.

2.2 Multiplicative Updating Algorithm to Solve Parameters of SVDD Model

The general formulation of the nonnegative quadratic programming is analyzed firstly. Consider the following minimization problem of quadratic function which contains inequality constraints

$$\min F(X) = \frac{1}{2} X^T \mathbf{A} X + \mathbf{b} X \tag{9}$$

In Eq.(9), the i th component of X is denoted as X_i, and the constrain conditions are $X_i \ge 0, \ i = 1, 2, ..., N.$ And $\mathbf{A}$ is $N \times N$ symmetric nonnegative matrix. Iterative algorithm can be constructed to solve the minimum value of Eq.(9) with the nonnegative constrain condition of X shown as reference[19]. Matrix $\mathbf{A}$ is expressed by the subtraction of two nonnegative matrix $\mathbf{A}^+$ and $\mathbf{A}^-$ as

$$\mathbf{A} = \mathbf{A}^+ - \mathbf{A}^- \tag{10}$$

where

$$(\mathbf{A}^+)_{i,j} = \begin{cases} (\mathbf{A})_{i,j} \text{ if } (\mathbf{A})_{i,j} > 0 \\ 0 \qquad \text{others} \end{cases}, \quad (\mathbf{A}^-)_{i,j} = \begin{cases} -(\mathbf{A})_{i,j} \text{ if } (\mathbf{A})_{i,j} < 0 \\ 0 \qquad \text{others} \end{cases} \tag{11}$$

and $(\mathbf{A})_{i,j}$ is the (i, j) component of matrix $\mathbf{A}$ in Eq.(11). The iterative formulation of the multiplicative updating algorithm is as [19]

$$X_i^{(k+1)} = X_i^{(k)}[\frac{-b_i + \sqrt{b_i^2 + 4(\mathbf{A}^+ X^{(k)})_i (\mathbf{A}^- X^{(k)})_i}}{2(\mathbf{A}^+ X^{(k)})_i}] \tag{12}$$

where $k \in \mathbf{Z}^+$ is iterative times, b_i, $(\mathbf{A}^+ X^{(k)})_i$ and $(\mathbf{A}^- X^{(k)})_i$ are the i th component of vector $\mathbf{b}$, $\mathbf{A}^+ X^{(k)}$ and $\mathbf{A}^- X^{(k)}$.

When the iterative multiplicative updating algorithm is used in the Lagrange optimization problem shown as Eq.(7), and let $(\mathbf{A})_{i,j} = x_i \cdot x_j$, $b_i = x_i \cdot x_i$, the iterative computing formulation for the parameters of the SVDD model is shown as

$$\alpha_i^{(k+1)} = \alpha_i^{(k)}[\frac{-b_i + \sqrt{b_i^2 + 4(\mathbf{A}^+ \alpha^{(k)})_i (\mathbf{A}^- \alpha^{(k)})_i}}{2(\mathbf{A}^+ \alpha^{(k)})_i}] \ . \tag{13}$$

2.3 The SVDD Model Containing Two Classes of Samples

The SVDD model with one-class samples can be extended to the case of samples with two classes. Consider dataset $\{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$ come from two different classes of samples, where N is the number of samples. And x_i is the feature vector of the i th sample, $y_i = 1$ or -1 , $i = 1, 2, ..., N$. Not losing generality, for the samples $x_i, i = 1, 2, ..., l$, let $y_i = 1$, and for the samples x_i, $i = l+1, l+2, ..., N$, let $y_i = -1$. In other words, $\{x_i, i = 1, 2, ..., l\}$ are the positive samples, and $\{x_i, i = l+1, l+2, ..., N\}$ are negative samples or outliers. The positive samples $\{x_1, x_2 ..., x_l\}$ are in the hyper-sphere, and the negative samples $\{x_{l+1}, x_{l+2} ..., x_N\}$ are outside the hyper-sphere in the SVDD model. Slack variable $\xi_i^+ \geq 0, \ (i = 1, 2, ..., l)$ and $\xi_i^- \geq 0, \ (i = l+1, \ l+2, ..., N)$ are introduced in the objective function for each sample in the dataset similar with in one-class case. The problem of minimizing the radius of the hyper-sphere can be formulated by the following quadratic programming with inequality constraints

$$\begin{cases} \min R^2 + C_1 \sum_{i=1}^{l} \xi_i^+ + C_2 \sum_{i=l+1}^{N} \xi_i^- \\ sub: \left\| x_i - \mathbf{a} \right\|^2 \leq R^2 + \xi_i^+, \xi_i^+ \geq 0, i = 1, 2, ..., l; \\ \qquad \left\| x_i - \mathbf{a} \right\|^2 \geq R^2 - \xi_i^-, \xi_i^- \geq 0, i = l+1, l+2, ..., N. \end{cases} \tag{14}$$

where the positive constant parameters C_1 and C_2 are penalty factors. Using Lagrange multiplier algorithm for Eq.(14), we can draw the corresponding Lagrange function as

$$
\begin{aligned}
&L(R,\mathbf{a},\alpha,\beta,\xi_i^+,\xi_i^-)= \\
&R^2+C_1\sum_{i=1}^{l}\xi_i^+ +C_2\sum_{i=l+1}^{N}\xi_i^- -\sum_{i=1}^{l}\beta_i\,\xi_i^+ -\sum_{i=l+1}^{N}\beta_i\,\xi_i^- \\
&-\sum_{i=1}^{l}\alpha_i\,(R^2+\xi_i^+ -\|x_i-\mathbf{a}\|^2)-\sum_{i=l+1}^{N}\alpha_i\,(\|x_i-\mathbf{a}\|^2+\xi_i^- -R^2)
\end{aligned}
\tag{15}
$$

where $\alpha_i \ge 0, \beta_i \ge 0$ are Lagrange multipliers. Similar with Eq.(3), Lagrange function L should be minimized with respect to $R,\mathbf{a},\xi_i^+,\xi_i^-$ and maximized with respect to α_i and β_i. After computing the extremum conditions of Lagrange function L, the dual form of the Lagrange optimization problem Eq.(15) are shown as following quadratic programming problem containing inequality constraints

$$
\begin{cases}
\max \sum_{i=1}^{l}\alpha_i(x_i\cdot x_i)-\sum_{i=l+1}^{N}\alpha_i(x_i\cdot x_i)- \\
\quad \sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j(x_i\cdot x_j)+2\sum_{i=1}^{l}\sum_{j=l+1}^{N}\alpha_i\alpha_j(x_i\cdot x_j)-\sum_{i=l+1}^{N}\sum_{j=l+1}^{N}\alpha_i\alpha_j(x_i\cdot x_j) \\
sub:\ \sum_{i=1}^{l}\alpha_i-\sum_{i=l+1}^{N}\alpha_i=1,\ \ 0\le\alpha_i\le C_1, i=1,2,...,l; \\
\quad 0\le\alpha_i\le C_2,\ \ i=l+1,l+2,...,N.
\end{cases}
\tag{16}
$$

Let $\alpha_i' = y_i\alpha_i$, then we have $\sum_{i=1}^{N}\alpha_i' = 1$ and $\mathbf{a}=\sum_{i=1}^{N}\alpha_i' x_i$. Then Eq.(16) can be simplified as

$$
\begin{cases}
\max \sum_{i=1}^{N}\alpha_i'(x_i\cdot x_i)-\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i'\alpha_j'(x_i\cdot x_j) \\
sub:\ \sum_{i=1}^{N}\alpha_i' = 1,\ 0\le\alpha_i\le C_1, i=1,2,...,l; \\
\qquad\qquad 0\le\alpha_i\le C_2, i=l+1,l+2,...,N.
\end{cases}
\tag{17}
$$

Similar with Eq.(7), the quadratic programming problem containing inequality constraints denoted as Eq.(17) can be solved using the iterative multiplicative updating algorithm. Then the parameters of the SVDD model Eq.(16) is

$$
\alpha_i = \frac{\alpha_i'}{y_i}, i=1,2,...,N..
\tag{18}
$$

Then the radius R and the center $\mathbf{a}$ of the hyper-sphere of the SVDD model are achieved. So the dataset containing two classes of samples are separated by the the

hyper-sphere. Decision-making function to classify a new sample using the hyper-sphere of the SVDD model will be discussed.

2.4 Decision-Making Function and Kernel Transformation in Feature Space

The SVDD model is achieved from the known samples. If a new sample is in the hyper-sphere of the SVDD model, it belongs to the positive class. Otherwise it is negative one or outlier. So the following decision-making function can be constructed

$$y(x) = \mathrm{sgn}(R^2 - ((x \cdot x) - 2\sum_{i=1}^{N} \alpha_i (x_i \cdot x) + \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j (x_i \cdot x_j))) . \quad (19)$$

For a new sample denoted by x, if the computing result of Eq.(19) is $y(x) \geq 0$, it belongs to the positive class. And if the result is $y(x) < 0$, it is outlier or negative sample.

In order to determine the decision-making function, the radius and the center of the hyper-sphere should be computed. From Eq.(8), we can see that most of parameters α_i are zero. Only part of parameters are non-zero. The samples corresponding the non-zero α_i values are called support vectors. They determine the radius and the center of the hyper-sphere. The center $\mathbf{a}$ is calculated by Eq.(5). We assume that $\alpha_k \neq 0$ for some support vector x_k. Then the radius of the hyper-sphere can be calculated as

$$R = ((x_k \cdot x_k) - 2\sum_{i=1}^{N} \alpha_i (x_k \cdot x_k) + \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j (x_i \cdot x_j)))^{\frac{1}{2}} . \quad (20)$$

If the inner product of x_i and x_j is substituted by kernel function $K(x_i, x_j)$, the decision-making function can be shown as

$$y(x) = \mathrm{sgn}(R^2 - (K(x,x) - 2\sum_{i=1}^{N} \alpha_i K(x_i, x) + \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j K(x_i, x_j))) . \quad (21)$$

Kernel functions are usually constructed by mapping functions from the sample space to the high dimensional feature space. Examples of the kernel functions are linear kernel, polynomial kernel and the Gaussian kernel and so on.

3 Predication Method of Customer Credit Scoring Based on the SVDD Model

General predication methods driven by known samples will be discussed briefly. Then new predication method of customer credit scoring of banks based on the SVDD classification model is proposed. Merits of the proposed method over the SVM model based method are analyzed.

There are mainly two parts called as learning stage and predicating stage in pattern predication methods driven by dataset of samples. The processes of predication method are as following. Firstly, the original dataset are collected, cleared, completed and normalized in preprocessing stage. The normalized dataset $\{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$ are established, where $x_i \in R^p$ is the input p -dimension feature vector which represents known sample, and N is the number of samples. $y_i \in \{+1, -1\}$ is the decision value corresponding the input vector x_i , $i = 1, 2, ..., N$. In the learning stage, suitable mathematical models are selected to establish the classification decision- making function $y = f(x, P)$, where P is parameter set of the models. Different models such as artificial neural networks[8], SVM model[6], and SVDD model can be used to simulate the decision-making function. After establishing the models, the optimized decision-making function $y = f(x, P^*)$ will be achieved by some learning algorithm using the dataset of the known samples, where P^* denoted the optimized parameter set. At last, the predication result for a new sample x^* can be gained by inputting the sample x^* into the optimized decision-making function and computing the value of the function $y = f(x^*, P^*)$ in the predicating stage.

When the SVDD classification model is used in predication of customer credit scoring of banks, positive samples (well-recorded customers) are assumed to be in the hyper-sphere, and negative samples (bad-recorded customers) be out of the hyper-sphere. The predication method of customer credit scoring of banks based on the SVDD model can be summed as following. The dataset of learning samples $\{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$ are achieved after the preprocessing stage from existing original data records. Some special kernel function is selected, and the SVDD classification model is determined. Parameters $\{\alpha_i, i = 1, 2, ..., N\}$ of the SVDD model denoted by the quadratic programming problem Eq.(16) are solved by using the iterative multiplicative updating algorithm Eq.(13). Then the center $\mathbf{a}$ and the radius R of the hyper-sphere are calculated by Eq.(5) and Eq. (20) separately. Such, the SVDD classification model is established by the known samples. In the predication stage, when a new sample x^* to be predicated is substituted into the decision-making function Eq.(21), comes out the predicating value. We can explain which class the sample belongs to according to the value meaning.

The difference of the SVDD classification model and the SVM model will be discussed briefly. The SVM model mainly classifies samples of two or more classes, the principle of which is to maximize the margin hyper-plane that gives the maximum separation between two classes of samples. Both positive and negative classes of samples are needed in the processes of computing the parameters of the model. But in the SVDD model, optimal hyper-sphere is established that contain the most of positive samples and exclude the most of outliers. So the model can be solved by one class of samples or two classes of samples. The SVDD classification model has following characteristic cs and advantages over the SVM model:

1) The support vectors in the SVM model are the learning samples that determine the margin of hyper-plane. The support vectors in the SVDD model are the samples that are located at the boundary of the hyper-sphere. Both experiments and theoretical analysis showed that the number of support vectors of the latter is smaller than the former. So the computational costs of the SVDD model are smaller than that of the SVM model usually [21].

2) The numbers of two classes of samples influence the decision-making function of the SVM model greatly. So some scholars determine the weights of positive and negative samples in weighted SVM model by using the ratio of the number of positive and negative samples[17]. But in practical applications, the ratio of the number of positive and negative samples may be unknown. So the ratio is determined subjectively. On the other hand, the decision-making function is established by the hyper-sphere which is solved by support vectors located at the boundary of the hyper-sphere in the SVDD model. The ratio of the number of positive and negative samples has less influence in the decision-making function because the hyper-sphere can be determined by only one class of samples in some extreme situations. Simulation results in next section illustrate this advantage.

3) There are many matrix operations in solving the parameters denoted by quadratic programming problems in the SVM model, which take up much time [22]. The iterative multiplicative updating algorithm is used to solve the parameters of the SVDD model in our work. Not only the solving processes are simplified but also the computing efficiency is improved. Following experiments on artificial synthesized dataset and benchmark credit dataset of banks show the improvement of the proposed credit scoring predication method.

4 Experimental Results

Experiments on artificial dataset and benchmark individual credit dataset of banks are performed in this section, and effectiveness of the proposed customer credit scoring method based on the SVDD classification model with imbalanced dataset is illustrated. The weighted LS-SVM model is a new powerful classification model for imbalanced datasets proposed recently in machine learning [12]. So the proposed predication method is compared with the weighted LS-SVM model based method using same samples come from imbalanced datasets. We mainly want to indicate the improvement of the learning and predicating processes of the proposed method for the imbalanced dataset which containing different numbers of positive and negative samples. The learning and predicating accuracies are compared using the proposed method and the weighted LS-SVM model.

4.1 Experiments on Artificial Dataset

Two-spiral classification is one of classical problems in pattern recognition[12]. The distributions of the samples of two-spiral function are shown as figure one.

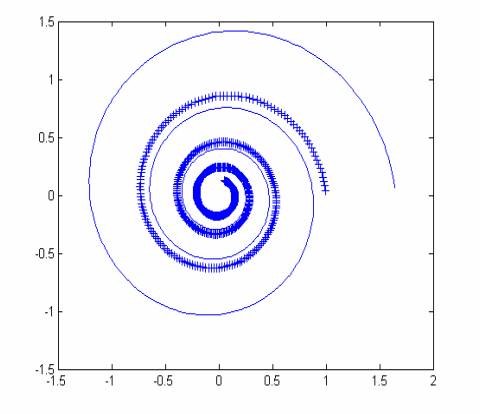

Fig. 1. The distribution of the samples of two-spiral function

Samples of the first class in figure 1 are produced by Eq.(22).

$$\begin{cases} x_1(t) = \exp(-0.2t)\cos(2t) \\ y_1(t) = \exp(-0.2t)sin(2t) \end{cases}. \tag{22}$$

And samples of the second class are produced by Eq.(23).

$$\begin{cases} x_2(t) = \exp(-0.2t + 0.5)\cos(2t) \\ y_2(t) = \exp(-0.2t + 0.5)sin(2t) \end{cases} \tag{23}$$

The values of parameter t are from 0.02 to 10, and the step is 0.02. There are five hundreds samples in each class. The first class is denoted as positive sample set, and the second is negative. The objective of experiments is to illustrate the improvement of accuracy of learning and predicating of SVDD model over the SVM model for classification problems with imbalanced dataset. The learning and predicating experiments using different number of samples are performed. And Gauss function Eq.(24) is used as kernel function.

$$K(x, y) = \exp(-\frac{\|x - y\|^2}{2\delta^2}) \tag{24}$$

Table one shows a group of learning and predicating results of two-spiral function using both models with several different numbers of positive and negative samples. This group of experiments is designed as follows. The parameter δ of the kernel function Eq.(24) is 2, and penalty factors C_1 and C_2 are both 4 in Eq.(17).

Table 1. Learning and predicating results of two-spiral classification problem

Model	WLS-SVM		SVDD	
(n^+, n^-)	ARL(%)	ARP(%)	ARL(%)	ARP(%)
(200,200)	99.25	96.33	98.25	95.83
(200,180)	95.26	94.52	96.32	94.19
(200,160)	92.78	91.72	94.72	93.28
(200,140)	92.65	91.67	96.47	92.88
(200,120)	90.00	88.82	94.06	92.35
(200,100)	85.67	87.43	93.33	90.86
(200, 80)	87.86	84.86	93.57	90.56
(200, 60)	85.39	84.46	91.92	91.62

In table one, n^+ and n^- denote the numbers of positive and negative samples. In learning process, n^+ positive and n^- negative samples are selected randomly from artificial two-spiral dataset show in table one. And same learning samples are used in both models in the each group of comparing experiment. In predicating process, residual parts of the artificial dataset are as samples to be predicated. In table one, (n^+, n^-) denotes numbers of learning samples, ARL denotes accuracy rate of learning, and ARP denotes accuracy rate of predicting. Accuracy rate of learning is the ratio of correctly classified samples with the total learning samples. Accuracy rate of predicting is the ratio of correctly classified samples with the total predicated samples. All experiments are performed under following conditions: hardware CPU Pentium4 2.4 GHZ, RAM 512MB; software Windows XP and Matlab7.0.

From the experimental results shown in table one, we can see the WLS-SVM model has higher accuracy of learning and predicating than the SVDD model slightly when the number of positive samples equals to the number of negative. But when the numbers of positive samples are different from the number of negative greatly, the SVDD model keeps higher accuracy of learning and predicating compared with the WLS-SVM model apparently. So the SVDD model is more effective to deal with classification problems with imbalanced artificial dataset.

4.2 Experiments on Benchmark Dataset

In this subsection, the experiment samples are selected from an opening database of computer institute of UCI university [23]. There are three sets of individual credit scoring of banks about Australia, Germen, and Japan in the database. Experimental results using the database about Australia are listed following. There are total 690 samples in the database. The number of positive sample (good credit) is 307, and others are negative sample (bad credit). There are fourteen index of evaluation and one credit value to compose a sample.

The database is preprocessed firstly. All the records in the database have been numerically disposed. So the index and credit value are expressed by numerical values accordingly. We can see that the input vector which denotes each sample is a fourteen-dimension vector from the index of credit of each sample. The inner product of two vectors is relative with each component of the both vectors. But there exists magnificent discrimination in the value ranges of each index in original database. In order to balance the effect of each component of the input vector (each credit index), all the values of index are normalized using as

$$\hat{x}_{i,j} = \frac{x_{i,j} - \min x_j}{\max x_j - \min x_j} \tag{25}$$

where $\max x_j$ and $\min x_j$ denote the maximum and minimum values of the j th index of all the samples in the database, and $j = 1, 2, ..., 14$, $i = 1, 2, ..., 690$. Such, the normalized data are achieved.

Table 2. Learning and predicating results of individual credit scoring of banks

Model	**WLS-SVM**		**SVDD**	
(n^+, n^-)	ARL(%)	ARP(%)	ARL(%)	ARP(%)
(200,150)	73.71	62.35	80.29	70.29
(200,160)	75.83	63.33	81.94	68.49
(200,170)	79.46	68.13	83.24	71.88
(200,180)	81.84	66.67	85.26	72.58
(200,190)	84.10	69.67	83.33	73.67
(200,200)	85.75	71. 38	85.00	74.48
(190,200)	82.82	70.33	84.87	72.33
(180, 200)	82.89	67.74	84.47	73.55
(170, 200)	78.11	66.56	83.51	73.44
(160, 200)	76.67	64.85	80.56	70.30
(150, 200)	75.72	64.41	78.86	68.82

Table two shows a group of learning and predicating results of individual credit scoring of banks using the dataset of Australia using the weighted LS-SVM model and the SVDD model with different numbers of positive and negative samples. The kernel function and parameters are same as experiments on the artificial dataset. In table two, n^+ and n^- denote the numbers of positive and negative samples. In learning process, n^+ positive and n^- negative learning samples are selected randomly

from the database about Australia. And same learning samples are used in both models in the each group of comparing experiment. In predicating process, residual parts of the dataset are as samples to be predicated. Experiment environments are same as those of the artificial dataset.

From table two, we can see that the SVDD model and the WLS-SVM model have similar accuracy of learning when the number of positive samples equals to the number of negative ones. The SVDD model has higher accuracy of learning and predicating compared with the WLS-SVM models when the number of positive samples are different from the number of negative ones greatly similar with experimental results on the artificial dataset. We see that the accuracies of learning and predicating using the dataset of the individual credit scoring of Australia are not as well as those of the artificial dataset of two-spiral. One reason is that the dataset of the individual credit scoring is not separable strictly itself [24].

5 Conclusions

New customer credit scoring predication method based on the SVDD classification model with imbalanced dataset was proposed in this paper. Main predication methods were reviewed firstly. Then the SVDD classification model for imbalanced dataset was analyzed. And the multiplicative updating principle to solve the parameters of the model was discussed. Following, new learning and predicating method of customer credit scoring of banks based on the SVDD model was proposed. At last, Experiments on the synthesized two-spiral dataset and the benchmark dataset of individual credit of banks using the proposed method and the WLS-SVM-based method were performed. Experimental results illustrated that the proposed method is more effective than the WLS-SVM-based method for classification problems with imbalanced dataset such as predication of customer credit scoring of banks. The learning and predicating accuracies of SVDD model are higher than the LS-SVM model under same experiment conditions.

Acknowledgment

The work was partly supported by the National Natural Science Foundation of China (70971083), Leading Academic Discipline Program, 211 Project for Shanghai University of Finance and Economics(the 3rd phase) and the National Post-Doctoral Foundation of China(20080440644).

References

1. Nanda, S., Pendharkar, P.C.: Development and Comparison of Analytical Techniques for Predicting Insolvency Risk. International Journal of Intelligent Systems in Accounting Finance and Management 10(3), 155–168 (2001)
2. Kay, S.M.: Fundamentals of Statistical Signal Processing. Estimation Theory, Detection Theory, vol. I, II. Prentice-Hall, New Jersey (1998)
3. Altman, E.L.: Financial Ratios Discriminate Analysis and the Prediction of Corporate Bankruptcy. Journal of Finance 23(3), 589–609 (1968)

4. Pendharkar, P.C.: A Computational Study on the Performance of ANNs under Changing Structural Design and Data Distributions. European Journal of Operational Research 138(1), 155–177 (2002)
5. Gomez Skarmeta, A.F., Delgado, M., Vila, M.A.: About the Use of Fuzzy Clustering Techniques for Fuzzy Model Identification. Fuzzy Sets and Systems 106(2), 179–188 (1999)
6. Bernhard, S., Sung, K.K.: Comparing Support Vector Machines with Gaussian Kernels to Radical Basis Function Classifiers. IEEE Transaction on Signal Processing 45(11), 2758–2765 (1997)
7. Thomas, L.C., Edelman, D.B., Crook, J.N.: Credit Scoring and Its Applications. Cambridge University Press, Cambridge (2002)
8. Desai, V.S., Crook, J.N., Overstreet, J.G.A.: A Comparison of Neural networks and Linear Scoring Models in the Credit Union Environment. European Journal of Operational Research 95(1), 24–37 (1996)
9. David, R.H., Edelman, D.B., Gammerman, A.J.: Machine Learning Algorithm for Credit-card Applications. Journal of Mathematics Applied in Business and Industry 4(1), 43–51 (1992)
10. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (1998)
11. Cortes, C., Vapnik, V.: Support Vector Networks. Machine Learning 20(3), 273–297 (1995)
12. Suykens, J.A.K., Vandewalle, J.: Least Squares Support Vector Machine Classifiers. Neural Proc. Lett. 9(3), 293–300 (1999)
13. Zhang, L., Zhou, W., Jiao, L.: Wavelet Support Vector Machine. IEEE Trans. on Systems, Man, and Cybernetics-part B: Cybernetics 34(1), 34–39 (2004)
14. Tax, D.M.J., Duin, R.P.W.: Support Vector Data Description. Machine Learning 54(1), 45–66 (2004)
15. Wang, Y., Wang, S., Lai, K.K.: A New Fuzzy Support Vector Machine to Evaluate Credit Risk. IEEE Transaction on Fuzzy Systems 13(6), 820–831 (2005)
16. Zadrozny, B., Langford, J., Abe, N.: Cost Sensitive Learning by Cost Proportionate Example Weighting. In: The 3rd IEEE International Conference on Data Mining, pp. 435–442 (2003)
17. Lin, Y., Lee, Y., Wahba, G.: Support Vector Machines for Classification in Nonstandard Situations. Machine Learning 46(2), 191–202 (2002)
18. Thomas, L.C.: A Survey of Credit and Behavioral Scoring: Forecasting Financial Risk of Lending to Customers. International Journal of Forecasting 16(2), 149–172 (2002)
19. Tax, D.M.J.: One-class Classification. PhD thesis, Delft University of Technology, http://www.ph.tn.tudelft.nl/~davidt/thesis.pdf (last accessed, March 2010)
20. Sha, F., Lin, Y., Saul, L.K., Lee, D.D.: Multiplicative Updates for Nonnegative Quadratic Programming. Neural Computation 19(8), 2004–2031 (2007)
21. Tax, D.M.J., Duin, R.P.W.: Support Vector Domain Description. Pattern Recognition Letters 20(11-13), 1191–1199 (1999)
22. Sha, F., Saul, L.K., Lee, D.D.: Multiplicative Updates for Nonnegative Quadratic Programming in Support Vector Machines. In: Becker, S., Thrun, S., Obermayer, K. (eds.) Advances in Neural Information Processing Systems, vol. 15, pp. 897–904 (2003)
23. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository, http://www.ics.uci.edu/~mlearn/MLRepository.html (last accessed, March 2010)
24. Schölkopf, B., Smola, A.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge (2002)