Journal of Management Science and Statistical Decision
Vol.7 No1.(2010) p110-117

Twin-SVDD Classifier with the Conception of Relative Distance

Lei Yang and Wei-Min Ma
School of Economics and Management
Tongji University
1239 Si-Ping Road, Shanghai, China
yangyoungya@sina.com; mawm@buaa.edu.cn

Bo Tian
School of Information Management & Engineering
Shanghai University of Finance Economic
777 Guo-ding Road, Shanghai, China
youngtb@sina.com

***Abstract*—Twin support vector domain description(Twin-SVDD) classifier with the conception of relative distance was proposed in this paper. The preliminary SVDD model described the target dataset with one class by constructing a compact optimized hypersphere in feature space. And the model was effective to deal with problems of pattern classification with imbalanced dataset such as outlier detection. But only information about the positive class of dataset was used in the preliminary SVDD model. As for binary classification problem, inspired by the construction of Twin support vector machine where nonparallel planes were solved separately, the Twin-SVDD model was proposed. Two optimized hyperspheres which described positive and negative class of datasets were constructed separately in the Twin-SVDD model. So information about both classes of dataset was used. And then new classification decision-making function was constructed based on the parameters of the Twin-SVDD model with the conception of relative distance. At last, experiments were performed. Experimental results showed that the Twin-SVDD model was more effective than the preliminary SVDD model when dealing with pattern classification problems. And the proved classification decision-making function improved the performance of the Twin-SVDD model.**

Keywords-support vector domain description; twin support vector domain description; relative distance; pattern classification

I. INTRODUCTION

Classical methods of pattern classification mainly included parametric models such as statistical testing method, Bayes discriminate method, Fisher discriminate method, log-linear regression model and so on[1-2]. In classical methods, the numbers of samples were usually assumed to be sufficiently large, and samples were assumed to be some known distribution. But samples are usually finite even deficient in practice, and distributions of samples are even unknown. So non-parametric pattern classification methods such as neural networks, clustering method, support vector machine(SVM) were proposed in recent years, which were based on sample dataset[3-6]. Data-driven classification methods were based on statistical learning theories, and the disadvantages of statistical asymptotic theory can be tided over. Main principle of data-driven classification methods is to construct decision-making function by learning processes of the dataset of the objects with small or finite samples. Prior knowledge about the samples needed not to be known. Minimization of experimental risk is used in the neural networks predication method, which makes the total output error be minimized by adjusting the weights of the neural networks using some learning algorithms[3]. But several disadvantages of neural networks such as over-fitting phenomenon in learning processes, lack of generalization ability, and local extremum values limited their practical applications[7]. Cortes and Vapnik proposed SVM models by constructing a decision superplane that maximized the margin of two classes of samples, which minimized the structural risk [7-8]. The complexity of models and experience risk can be balanced effectively in the SVMs, and generalization ability of model was improved. Problems such as small number of samples, nonlinear map, high dimension description, and local extremum values can also be solved. So the SVMs are very suitable to be used in problems of pattern classification with small samples, approximation of functions and so on.

Then some improved SVM models were proposed by different researchers. Suykens proposed Least squares SVM[9]. Zhang proposed Wavelet SVM[10]. Doumpos proposed additive SVM[11]. Jayadeva proposed Twin SVM[12]. Tax proposed support vector domain description(SVDD) model[13]. Applications of the SVM were also investigated. The principle of SVDD models was to construct a hypersphere with minimized the radius which contains the most of positive examples, and others samples named outliers were located outside of the hypersphere[13-16]. So computing tasks of SVDD models were to calculate the radius and center of the hypersphere using the given samples. The SVDD models based on data description method were mainly used to deal with the problem of one-class classification such as to describe dataset and detect outliers[13-14]. The one-class dataset was described by using the samples located at boundary of hypersphere in the SVDD models, which were named support vectors. In the preliminary SVDD model, dataset was described by an optimized hypersphere, and decision-making function was based on the hypersphere. Because only information about positive class of samples was used in the preliminary SVDD model, the model may be not the best. On the other hand, information about negative class of sample dataset was available. So in this paper, we proposed an improved SVDD model to solve pattern classification problem. Inspired by the Twin SVM model proposed Jayadeva where two hyperplanes need not be parallel as in SVM model[12,17], we proposed the Twin-SVDD model which constructed two optimized hyperspheres to describe each class of positive and negative samples separately and new classification decision-making function was established based on the parameters of the Twin-SVDD model with the conception of relative distance. The paper was organized as following. Main classes of pattern classification methods were reviewed and the spirit of this paper was discussed in section one. The SVDD model to describe one class of samples was introduced in section two. The preliminary SVDD classification model was analyzed in section three. The Twin-SVDD model was proposed in section four. Then new decision-making function which was based on the parameters of the Twin-SVDD model with the conception of relative distance was discussed in section five. Experimental results of the proposed the Twin-SVDD model and the preliminary SVDD model were reported in section six. And conclusions were drawn in the last section.

II. THE SVDD MODEL TO DESCRIBE ONE CLASS OF SAMPLES

A compact high-dimensional hypersphere with minimized the radius is established in the SVDD model. The positive examples are included in the hypersphere, and outliers are located outside of the hypersphere. So the SVDD model can be used in describing the target dataset or detecting outliers. Following the preliminary SVDD model is reviewed[13-14].

The objective of SVDD model is to describe the dataset by using the hypersphere with minimized radius R in feature space. In other words, the target samples are located in an optimized hypersphere. Let dataset $\{x_i, i=1,2,...N\}$ be training samples. The mathematic form of the model is minimizing the function $F(R,\mathbf{a})=R^2$ with the constraint condition $\|x_i-\mathbf{a}\|^2 \le R^2$, $(\forall i=1,2,...N)$. Thinking the influence of outliers or noise, distances from samples $\{x_i, i=1,2,...N\}$ to the center $\mathbf{a}$ of the hypersphere are not strictly smaller than R. But large distance should be penalized. 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 hypersphere can be shown as 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 penalty factor. It controls the trade-off between the radius of hypersphere and the error. Using Lagrange multipliers algorithm for Eq.(1), the corresponding Lagrange function is

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

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 . 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) due to $\alpha_i \ge 0, \beta_i \ge 0$. When Eq.(4-6) are substituted into Lagrange function Eq.(2), arrive at the dual form of the Lagrange optimization problem as

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

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

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

III. THE PRELIMINARY SVDD CLASSIFIER FOR TWO CLASSES OF SAMPLES

The SVDD model to describe one class of samples can be used in classification of samples of 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 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 positive samples, and $\{x_i, i=l+1,l+2,...,N\}$ are negative samples or outliers. In the SVDD models, we assume that positive samples $\{x_1,x_2...,x_l\}$ are located in the hypersphere, and negative samples $\{x_{l+1},x_{l+2}...,x_N\}$ are outside of the hyper-sphere. Slack variable $\xi_i^+ \ge 0,\ (i=1,2,...,l);\ \ \xi_i^- \ge 0,\ (i=l+1,$

$l+2,...,N)$ are introduced in the objective function for each sample of both kinds of data similar with in one-class model. The problem of minimizing the radius of the hypersphere can be formulated by the following quadratic programming with inequality constraints

$$\begin{cases} \min R^2 + C_1\sum_{i=1}^{l}\xi_i^+ + C_2\sum_{i=l+1}^{N}\xi_i^- \\ sub: \|x_i - \mathbf{a}\|^2 \le R^2 + \xi_i^+, \xi_i^+ \ge 0, i = 1,2,...,l. \\ \qquad \|x_i - \mathbf{a}\|^2 \ge R^2 - \xi_i^-, \xi_i^- \ge 0, i = l+1, l+2,...,N. \end{cases} \tag{9}$$

Using Lagrange multipliers algorithm for Eq.(9), we can draw the corresponding Lagrange function as

$$\begin{aligned} L(R,\mathbf{a},\alpha,\beta,\xi_i^+,\xi_i^-) &= R^2 \\ &+ C_1\sum_{i=1}^{l}\xi_i^+ + C_2\sum_{i=l+1}^{N}\xi_i^- - \sum_{i=1}^{l}\beta_i\,\xi_i^+ - \sum_{i=l+1}^{N}\beta_i\,\xi_i^- \\ &- \sum_{i=1}^{l}\alpha_i\,(R^2 + \xi_i^+ - \|x_i - \mathbf{a}\|^2) \\ &- \sum_{i=l+1}^{N}\alpha_i\,(\|x_i - \mathbf{a}\|^2 + \xi_i^- - R^2) \end{aligned} \tag{10}$$

where $\alpha_i \ge 0, \beta_i \ge 0$ are Lagrange multipliers. Similar with Eq.(2). 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, dual form of the Lagrange optimization problem Eq.(10) are shown as following quadratic programming problem with inequality constraints

$$\begin{cases} \max \sum_{i=1}^{l}\alpha_i(x_i\cdot x_i) - \sum_{i=l+1}^{N}\alpha_i(x_i\cdot x_i) \\ \quad -\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j(x_i\cdot x_j) + 2\sum_{i=1}^{l}\sum_{j=l+1}^{N}\alpha_i\alpha_j(x_i\cdot x_j) \\ \quad -\sum_{i=l+1}^{N}\sum_{j=l+1}^{N}\alpha_i\alpha_j(x_i\cdot x_j) \\ sub: \sum_{i=1}^{l}\alpha_i - \sum_{i=l+1}^{N}\alpha_i = 1, \\ \quad 0 \le \alpha_i \le C_1, i = 1,2,...,l; \\ \quad 0 \le \alpha_i \le C_2, \quad i = l+1, l+2,...,N. \end{cases} \tag{11}$$

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.(11) can be simplified as

$$\begin{cases} \max \sum_{i=1}^{N}\alpha_i'(x_i\cdot x_i) - \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i'\alpha_j'(x_i\cdot x_j) \\ sub: \sum_{i=1}^{N}\alpha_i' = 1, \\ \quad 0 \le \alpha_i \le C_1, i = 1,2,...,l; \\ \quad 0 \le \alpha_i \le C_2, \quad i = l+1, l+2,...,N. \end{cases} \tag{12}$$

Similar with Eq.(7), the quadratic programming problem containing inequality constraints denoted as Eq.(12) can be solved using the iterative multiplicative updating algorithm easily[18]. Then the radius R and center $\mathbf{a}$ of the hypersphere are sovled. So the dataset containing two classes of samples are separated by the hypersphere. And decision-making function of SVDD to classify a new sample can be constructed by the hypersphere. If a new sample x is located in the hypersphere, it belongs to the positive class. Otherwise it belongs to the negative one. So the following decision-making function can be constructed as

$$\begin{aligned} y(x) &= \mathrm{sgn}(R^2 \\ &-((x\cdot x) - 2\sum_{i=1}^{N}\alpha_i(x_i\cdot x) + \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j(x_i\cdot x_j))) \end{aligned} \tag{13}$$

For a new sample denoted by x, if the computing result of Eq.(13) is $y(x) \ge 0$, 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, R and center $\mathbf{a}$ of the hypersphere should be computed. In application, most of parameters α_i are zero. Only part of parameters are non-zero. The samples corresponding the non-zero α_i values are support vector. They determine the radius R center $\mathbf{a}$ of the hypersphere. the center $\mathbf{a}$ is calculated by Eq.(5). We assume that $\alpha_k \ne 0$ for some support vector x_k. Then radius R of the hypersphere can be calculated as

$$R = ((x_k\cdot x_k) - 2\sum_{i=1}^{N}\alpha_i(x_k\cdot x_k) + \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j(x_i\cdot x_j)))^{\frac{1}{2}} \tag{14}$$

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

$$\begin{aligned} y(x) &= \mathrm{sgn}(R^2 \\ &-(K(x,x) - 2\sum_{i=1}^{N}\alpha_i K(x_i,x) + \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j K(x_i,x_j))) \end{aligned} \tag{15}$$

The kernel functions are usually constructed by mapping function which satisfied kernel conditions, and examples are linear kernel, polynomial kernel and Gauss kernel and so on[17].

IV. The Proposed Twin-SVDD Model

Similar with the preliminary SVDD model, sample 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 i th sample, $y_i = +1$ or $-1, i = 1, 2, ..., N$. For samples $x_i, i = 1, 2, ..., l$, let $y_i = +1$, and for samples $x_i, i = l+1,$, $l+2, ..., N$, let $y_i = -1$. One optimized hypersphere was constructed to describe positive sample dataset $\{x_i, i = 1, 2, ..., l\}$, and another hypersphere was constructed to describe negative sample dataset $\{x_i, i = l+1,\ l+2, ..., N\}$ separately. Slack variable $\xi_i^+ \geq 0,\ (i = 1, 2, ..., l)$ and $\xi_i^- \geq 0,\ (i = l+1, l+2, ..., N)$ are introduced in objective function for each sample of both kinds of dataset similar with in the preliminary SVDD model because not all samples were located in both hyperspheres strictly. The problem of minimizing radius of the both hyperspheres of the Twin-SVDD model can be formulated by the following quadratic programming with inequality constraints

$$\begin{cases} \min R_1^2 + R_2^2 + C_1 \sum_{i=1}^{l} \xi_i^+ + C_2 \sum_{i=l+1}^{N} \xi_i^- \\ sub: \|x_i - \mathbf{a}_1\|^2 \leq R_1^2 + \xi_i^+, \\ \qquad \xi_i^+ \geq 0, i = 1, 2, ..., l; \\ \|x_i - \mathbf{a}_2\|^2 \leq R_2^2 + \xi_i^-, \\ \qquad \xi_i^- \geq 0, i = l+1, l+2, ..., N. \end{cases} \tag{16}$$

where positive constant parameters C_1 and C_2 are penalty factors. Using Lagrange multipliers algorithm for Eq.(16), we can draw the corresponding Lagrange function as

$$\begin{aligned} L(R_1, R_2, \mathbf{a}_1, \mathbf{a}_2, \alpha, \beta, \xi_i^+, \xi_i^-) &= R_1^2 + R_2^2 \\ &+ C_1 \sum_{i=1}^{l} \xi_i^+ + C_2 \sum_{i=l+1}^{N} \xi_i^- - \sum_{i=1}^{l} \beta_i\, \xi_i^+ - \sum_{i=l+1}^{N} \beta_i\, \xi_i^- \\ &- \sum_{i=1}^{l} \alpha_i\, (R_1^2 + \xi_i^+ - \|x_i - \mathbf{a}_1\|^2) \\ &- \sum_{i=l+1}^{N} \alpha_i\, (R_2^2 + \xi_i^- - \|x_i - \mathbf{a}_2\|^2) \end{aligned} \tag{17}$$

where $\alpha_i \geq 0, \beta_i \geq 0$ are Lagrange multipliers. Lagrange function L should be minimized with respect to $R_1, R_2,$ $\mathbf{a}_1, \mathbf{a}_2,\ \xi_i^+, \xi_i^-$, and maximized with respect to α_i and β_i . Set partial derivatives $R_1, R_2, \mathbf{a}_1, \mathbf{a}_2, \xi_i^+, \xi_i^-$ of Lagrange function L to be zero give the following formula

$$\frac{\partial L}{\partial R_1} = 0 \to \sum_{i=1}^{l} \alpha_i = 1 \tag{18}$$

$$\frac{\partial L}{\partial R_2} = 0 \to \sum_{i=i+1}^{N} \alpha_i = 1 \tag{19}$$

$$\frac{\partial L}{\partial \mathbf{a}_1} = 0 \to \mathbf{a}_1 = \sum_{i=1}^{l} \alpha_i x_i \tag{20}$$

$$\frac{\partial L}{\partial \mathbf{a}_2} = 0 \to \mathbf{a}_2 = \sum_{i=l+1}^{N} \alpha_i x_i \tag{21}$$

$$\frac{\partial L}{\partial \xi_i^+} = 0 \to C_1 - \alpha_i - \beta_i = 0, i = 1, 2, ..., l. \tag{22}$$

$$\frac{\partial L}{\partial \xi_i^-} = 0 \to C_2 - \alpha_i - \beta_i = 0, i = l+1, l+2, ..., N. \tag{23}$$

Resubstituting Eq.(18-23) into Eq.(17) results in

$$\begin{aligned} L &= \sum_{i=1}^{l} \alpha_i (x_i \cdot x_i) + \sum_{i=l+1}^{N} \alpha_i (x_i \cdot x_i) \\ &- \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j (x_i \cdot x_j) - \sum_{i=l+1}^{N} \sum_{j=l+1}^{N} \alpha_i \alpha_j (x_i \cdot x_j) \end{aligned} \tag{24}$$

And the dual form of the Lagrange optimization problem Eq.(16) are shown as following quadratic programming problem with inequality constraints

$$\begin{cases} \max \sum_{i=1}^{l} \alpha_i (x_i \cdot x_i) + \sum_{i=l+1}^{N} \alpha_i (x_i \cdot x_i) \\ \quad - \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j (x_i \cdot x_j) - \sum_{i=l+1}^{N} \sum_{j=l+1}^{N} \alpha_i \alpha_j (x_i \cdot x_j) \\ sub: \sum_{i=1}^{l} \alpha_i = 1,\ \sum_{i=l+1}^{N} \alpha_i = 1, \\ \quad 0 \leq \alpha_i \leq C_1, i = 1, 2, ..., l; \\ \quad 0 \leq \alpha_i \leq C_2,\ \ i = l+1, l+2, ..., N. \end{cases} \tag{25}$$

We noticed that parameters set $\{\alpha_i, i = 1, 2, ..., l\}$ and $\{\alpha_i, i = l+1, 2, ..., N\}$ can be solved separately in Eq.(25). The optimization problem Eq.(25) can be separated into two single SVDD with one-class samples such as

$$\begin{cases} \max \sum_{i=1}^{l} \alpha_i (x_i \cdot x_i) - \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j (x_i \cdot x_j) \\ sub: \sum_{i=1}^{l} \alpha_i = 1,\ 0 \leq \alpha_i \leq C_1, i = 1, 2, ..., l. \end{cases} \tag{26}$$

and

$$\begin{cases} \max \sum_{i=l+1}^{N} \alpha_i (x_i \cdot x_i) - \sum_{i=l+1}^{N} \sum_{j=l+1}^{N} \alpha_i \alpha_j (x_i \cdot x_j) \\ sub: \sum_{i=l+1}^{N} \alpha_i = 1,\ 0 \leq \alpha_i \leq C_2, i = l+1, l+2, ..., N. \end{cases} \tag{27}$$

After computing parameters set $\{\alpha_i, i = 1, 2, ..., l\}$ and $\{\alpha_i, i = l+1, 2, ..., N\}$, radius R_1, R_2 and centers $\mathbf{a}_1, \mathbf{a}_2$ of the two hyperspheres to describe positive or negative sample dataset of Twin-SVDD model can be calculated similarly with in the preliminary SVDD model. For example, let non-zero α_k belongs to the parameters set $\{\alpha_i, i = 1, 2, ...l\}$ for some support vector x_k, R_1 and the centers $\mathbf{a}_1$ can be calculated by Eq.(14) and Eq.(20).

V. Decision-making Function Based on The Conception of Relative Distance

Eq.(7), Eq.(12), Eq.(26) and Eq.(27) are all some kind of quadratic programming problem with inequality constraints. These quadratic programming problem can be solved by many existing algorithms such as multiplicative updating algorithm, SOM and so on[18-19]. After the radius R and center $\mathbf{a}$ of the optimal hypersphere were solved in preliminary SVDD, we can choose that positive samples were assumed to be in the hypersphere of feature space, and negative sample to be out of the hypersphere when the models were used in classification prediction. The classification decision-making function is Eq.(15). We know, there is only one hypersphere that should be constructed in the preliminary SVDD model. A new sample must lie in or out of the hypersphere surely. But two hyperspheres will be constructed in the Twin SVDD model. So the decision-making function may be more complex in the Twin SVDD model than the decision function Eq.(15) in the preliminary SVDD model. Following we construct a new decision-making function for the Twin SVDD model based on the conception of relative distance.

After training dataset of samples containing two classes, two hyperspheres in feature space determined by $\{\mathbf{a}_1, R_1\}$ and $\{\mathbf{a}_2, R_2\}$ were known in the Twin SVDD model. Three are three possible cases that a new sample x^* may encounter. There is non hypersphere that x^* is located in. There is one hypersphere that x^* is located in. And there are two hyperspheres that x^* is located in. The case that the sample x^* encounters can be decided using the following criterion

$$\left\| x^* - \mathbf{a}_i \right\| \le R_i, i = 1, 2. \tag{28}$$

If there is no inequality in Eq.(28) that that the sample x^* satisfied, comes the first case. According the decision in the preliminary SVDD model, there is no class that the sample belongs to. But we want to know which class the sample more likely belongs to. In our method, the relative distances from the sample x^* to the two hyperspheres were calculated by normalizing the distances firstly. Then the class is determined by comparing the relative distances from the sample x^* and the two hyperspheres. Above thought can be written as following: If

$$\frac{\left\| x^* - \mathbf{a}_1 \right\|}{R_1} \le \frac{\left\| x^* - \mathbf{a}_2 \right\|}{R_2} \tag{29}$$

the sample x^* belongs to class one, and vise versa. The reasonability normalized relative distance can be illustrated as figure 1. Figure 1(a) shows a sample x^* and two hyperspheres in two dimension space. Which hypersphere the sample belongs to more likely can not be detected straightly when the distances from the sample to the boundaries of the two hyperspheres are about equal shown as figure 1(a). When the distance from the sample x^* to the two hyperspheres is normalized, if $R_2 < R_1$ and the distances from the sample to the boundaries of the two hyperspheres are about equal, the sample x^* is more likely belongs to the first class shown as figure 1(b). In figure 1(b) $d_1 = \left\| x^* - \mathbf{a}_1 \right\| / R_1$, and $d_2 = \left\| x^* - \mathbf{a}_2 \right\| / R_2$. So in this case, the decision function is expressed as following: If Eq.(29) is satisfied, the sample x^* belongs to the first class, otherwise it belongs to the second class.

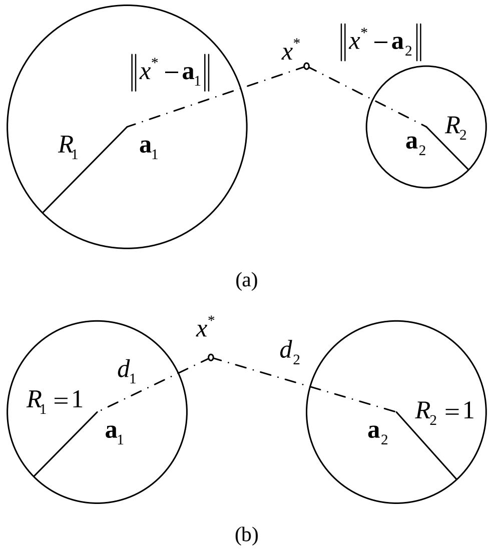

Figure 1. Sample x^* out of two hyperspheres $\{\mathbf{a}_1, R_1\}$ and $\{\mathbf{a}_2, R_2\}$ in two dimension space (a) and their relative location after normalized processing (b).

In the second case, if the sample x^* is located in the hypersphere $\{\mathbf{a}_i, R_i\}, (i = 1 \text{ or } 2)$, it belongs to that class. In the third case, if there are two hyperspheres that x^* is located in shown as figure 2(a), normalized processing was also adopted similar with in the first case. In figure 2(a), we can not detect which hypersphere the sample belongs to more likely when the distances from the sample to the centers of the two hyperspheres are about equal. Then the relative distances from the sample x^* to centers of the two hyperspheres are used, if $R_2 < R_1$ and the distances from the sample to the centers of the two hyperspheres are about equal, the sample x^* is more likely belongs to the first class similar with the first case shown in figure 2(b). d_1 and d_2 in figure 2(b) have same form as in figure 1(b). So if Eq.(29) is satisfied, the sample x^* also belongs to the first class.

We also noticed, in the second case, if the sample x^* is located in the hypersphere $\{\mathbf{a}_1, R_1\}$ and it belongs to the first class, because $\left\| x^* - \mathbf{a}_1 \right\| / R_1 \le 1$ and $\left\| x^* - \mathbf{a}_2 \right\| / R_2 > 1$, the decision function Eq. (29) is also satisfied naturally.

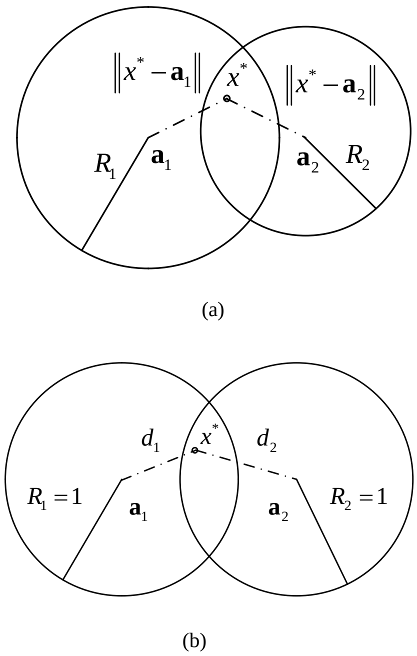

Figure 2. Sample x^* in two hyperspheres $\{\mathbf{a}_1, R_1\}$ and $\{\mathbf{a}_2, R_2\}$ in two dimension space (a) and their relative location after normalized processing (b).

Based on above analysis, the uniform decision-making function for a new sample x^* can be formulated as Eq.(29) in all three cases. And which case the sample x^* encounters need not to be distinguished any more. The distance expression after kernel transformation in feature space is given incidentally. The distance between x and y induced by inner product after kernel transformation can be expressed as

$$\|x - y\| = (K(x,x) - 2K(x,y) + K(y,y))^{\frac{1}{2}} \qquad (30)$$

VI. EXPERIMENTAL RESULTS

Effectiveness of the proposed Twin SVDD model was demonstrated experimentally by artificial and benchmark datasets in this section. The performances of proposed models and the preliminary SVDD model were compared. The experiments designs were presented, and then experimental results on both artificial and benchmark datasets were reported individually. We mainly want to indicate the improvement of the learning and predicating processes of the Twin SVDD model over the preliminary SVDD model under same experiment conditions.

A. Experiments Design

1) Data: Two-spiral classification is one of classical problems in pattern recognition. Artificial dataset was produced using two-spiral functions. The concrete method to produce samples and experimental results were given in next subsection. As for experiments on prediction classification of credit scoring, three credit dataset about Australian, Germen, and Japan from the UCI Machine Learning Repository were used[20]. Experiments on the Australian benchmark dataset were reported.

2) Models: Learning and predication processes of the preliminary SVDD Model Eq.(9) and the Twin-SVDD model Eq.(16) were compared. The concrete parameters should be given in advanced in these models, and they will be discussed along with experimental results.

3) Comparing criteria: We want to illustrate the improvement on the accuracy of learning and predicating of the proposed Twin SVDD model over the preliminary SVDD Model. In experiments, training error rate(%) and predicting error rate(%) were used to compare the performance. Because same numerical algorithm named iterative multiplicative updating algorithm was used to compute the parameters of the models in learning stages, and the dimensions of matrixes of the algorithm are the numbers of samples in experiments, the running times of algorithms were equal approximately. So the running times were not listed.

4) Setup: All experiments were performed under following conditions: hardware CPU Intel Core-2 -Duo 1.6GHZ, RAM 1024MB; software Windows XP and Matlab7.01.

B. Experiments on Synthesized Dataset

The samples of two-spiral are produced as following. Samples of the first class (marked as positive sample) were produced by

$$\begin{cases} x_1(t) = 10\exp(-0.02t)\cos(0.2t) \\ y_1(t) = 10\exp(-0.02t) sin(0.2t) \end{cases} \qquad (31)$$

and the samples of the second class (marked as negative sample) were produced by

$$\begin{cases} x_2(t) = 10\exp(-0.02t + 0.5)\cos(0.2t) \\ y_2(t) = 10\exp(-0.02t + 0.5) sin(0.2t) \end{cases} \qquad (32)$$

Values of parameter t were from 0.2 to 40, and the step was 0.2 in Eq.(31) and Eq.(32). There were two hundreds samples in each class. In experiments, the third polynomial function Eq.(33) and the Gauss function Eq.(34) were as kernel function. Both our experiments and references show that penalty factor parameter C (or C_1 and C_2) of SVDDs and parameter δ of the kernel function in Eq.(34) influence the accuracy of learning and predication for each model. Experiments of two models with different penalty factor parameter and different parameter δ of kernel function were performed.

$$K(x,y) = \|x^T y + 1\|^3 \qquad (33)$$

$$K(x,y) = \exp\left(-\frac{\|x - y\|^2}{\delta}\right) \qquad (34)$$

In our experiments, one hundred positive and one hundred negative samples selected randomly form the produced two-spiral dataset were as same training samples in two models. And samples from the residual dataset were used as predicated samples. We know if penalty factor parameter $C < 1/N$, two inequality constraints marked as Eq.(4-6) cannot be satisfied simultaneously. So parameters C should be a relative large positive constants satisfying $1/N < C \le 1$. Total $36(=\times 6\times 6)$ times experiments were performed. parameters C_1 and C_2 in the Twin-SVDD model were same as C in the preliminary SVDD Model. Parameters C and δ was set on a uniform coarse grid in the (C,δ) space for Gauss kernel function, and $C=[2^{-5},2^{-4},...,2^{0}]$, $\delta = [10^{-3},\ 10^{-2},...,10^{2}]$. We only list a group of experimental results. Table one showed a group of training error rate(TER) and predicting error rate(PER) of two-spiral dataset using the preliminary SVDD (P-SVDD) and Twin-SVDD (T-SVDD) by the third polynomial kernel function and the Gauss kernel function with when (C,δ) are $(0.25,1)$. In table one, TER was the ratio of numbers of misclassified samples to numbers of the total learning samples. PER was the ratio of numbers of misclassified samples to numbers of the total predicated samples.

TABLE I. TRAINING AND PREDICTING ERROR RATES OF TWO-SPIRAL CLASSIFICATION PROBLEM

Kernel Function	P-SVDD		T-SVDD	
	TER(%)	PER (%)	TER(%)	PER (%)
Polynomial	13.5	16.5	11.0	13.0
Gauss	11.0	13.5	10.5	12.0

Table one showed the Twin SVDD model has lower TER and PER than the preliminary SVDD model under same experiment conditions such as samples training dataset and parameters of models. From groups of experimental results we also found that the Twin SVDD model improved the performance of the preliminary SVDD model when experimental conditions are same.

C. *Experiments on benchmark dataset*

Series of experiments on learning and predicting of credit scoring were performed. Experiment samples comes from databases of computer institute of UCI university[20]. There are three sample sets of individual credit approval about Australian, Germen, and Japan in the database. Following report experimental results using the Australian benchmark dataset. There are total 690 samples in the database. Number of positive sample (good credit) is 307, and others are negative sample (bad credit). There are fourteen attribute index of credit and one credit value to compose a sample. The dataset were preprocessed firstly. All attribute names and values have been changed to numeric symbols to protect confidentiality of the samples. So the index and credit result were expressed by numeric values accordingly. We noticed that the input vector which denotes each sample is a fourteen-dimension vector from the index of each sample. The inner product or inner product after kernel transformation of two vectors were computed using each component of both vectors. But there existed magnificent discrimination in the value ranges of each credit 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 were normalized. Parameter C and δ in P-SVDD and parameters C_1 and C_2 in the T-SVDD were set similar with experiments on synthesized dataset. In training stage, two hundreds positive and two hundred negative same samples were selected in both models. And samples from the residual dataset were used as predicated samples. Table two showed a group of TER and PER of the credit approval database about Australian using P-SVDD and Twin-SVDD by the third polynomial kernel function and the Gauss kernel function with when (C,δ) are $(0.25,1)$.

TABLE II. TRAINING AND PREDICTING ERROR RATES OF THE CREDIT APPROVAL DATABASE ABOUT AUSTRALIAN

Kernel Function	P-SVDD		T-SVDD	
	TER(%)	PER (%)	TER(%)	PER (%)
Polynomial	24.50	28.46	22.50	24.62
Gauss	22.25	26.15	20.75	23.33

Representative results shown in table two illustrated that the Twin SVDD model has lower training error rate and predicting error rate compared with the preliminary SVDD models. We can see the accuracies of learning and predication for the real data of individual credit approval are not as well as those of the synthesized data of two-spiral. One reason is that the set of the real samples is not separable strictly. But similar with experiments on synthesized dataset, experimental results of learning and predicting the Australian credit approval database also showed that the Twin SVDD model improved the performance of the preliminary SVDD for pattern classification problems with real dataset under same experiments conditions such as samples dataset and parameters of models.

VII. CONCLUSIONS

Inspired by the Twin-SVM model proposed Jayadeva where two hyperplanes need not be parallel as in SVM model, the Twin-SVDDD classifier with relative distance conception was proposed in this paper. The preliminary SVDD model described target dataset with one-class by constructing a compact optimized hypersphere in feature space, and separated the positive samples with outliers by using the optimized hypersphere. But only information about the positive class of dataset was used in the preliminary SVDD model. As for binary classification problem, a new Twin-SVDD classifier that contains two hyperspheres was proposed in which two optimized hyperspheres described positive class of dataset and negative class of dataset separately. So information about both classes of dataset was used. And new classification decision-making function was constructed based on parameters of the

Twin-SVDD model with the conception of relative distance. Experimental results showed that the Twin-SVDD model has relative low training error rate and predicting error rate compared with the preliminary SVDD model when dealing with pattern classification problem under same experiments conditions such as samples dataset and parameters of models. We will extend our model to deal with multi-classification problems.

ACKNOWLEDGMENT

The work was partly supported by the National Post-Doctoral Foundation of China (20080440644), the National Natural Science Foundation of China(70671004), Program for New Century Excellent Talents in University (NCET-06-0172), A Foundation for the Author of National Excellent Doctoral Dissertation of PR China (200782), and the Shuguang Plan of Shanghai Education Development Foundation and Shanghai Education Committee (08SG21).

REFERENCES

[1] S. M. Kay. Fundamentals of statistical signal processing. Volume I: Estimation theory/Volume II: Detection theory[M]. Prentice-Hall, New Jersey, USA, 1998.

[2] S. Nanda, P. C. Pendharkar. Development and comparison of analytical techniques for predicting insolvency risk[J]. International Journal of Intelligent Systems in Accounting, Finance and Management, 2001, 10(3): 155-168.

[3] P. C. Pendharkar. A computational study on the performance of ANNs under changing structural design and data distributions[J]. European Journal of Operational Research, 2002, 138(1): 155-177.

[4] A. F. Gomez-Skarmeta, M. Delgado and M. A. Vila. About the use of fuzzy clustering techniques for fuzzy model identification[J]. Fuzzy Sets and Systems, 1999, 106(2): 179-188.

[5] S. Bernhard, K. K. Sung. Comparing support vector machines with Gaussian kernels to radical basis function classifiers[J]. IEEE Transaction on Signal Processing, 1997, 45(11): 2758-2765.

[6] D. Martens, J. Huysmans1, R. Setiono. Rule extraction from support vector machines: an overview of issues and application in credit scoring[J] Studies in Computational Intelligence,2008, 80: 33-63.

[7] V. Vapnik. The Nature of Statistical Learning Theory[M]. New York: Springer, 1998.

[8] C. Cortes, V. Vapnik. Support vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[9] J. A. K. Suykens, J. Vandewalle. Least squares support vector machine classifiers[J]. Neural Proc. Lett., 1999, 9(3): 293-300.

[10] L. Zhang, W. Zhou and L Jiao. Wavelet support vector machine[J]. IEEE Trans. on Systems, Man, and Cybernetics-part B: Cybernetics, 2004, 34(1): 34-39.

[11] M. Doumpos, C. Zopounidis and V. Golfinopoulou. Additive support vector machines for pattern classification[J]. IEEE Trans on Systems, Man, and Cybernetics-Part B: Cybernetics, 2007, 37(3): 540-550.

[12] Jayadeva, R. Khemchandani and S. Chandra. Twin support vector machines for pattern classification[J]. IEEE Trans on Pattern Anal. Machine Intell, 2007, 29 (5): 905-910.

[13] D. M. J. Tax, R. P. W. Duin. Support vector data description[J]. Machine Learning, 2004, 54(1): 45-66.

[14] D. M. J. Tax, R. P. W. Duin. Support vector domain description[J]. Pattern Recognition Letters, 1999, 20(11-13): 1191-1199.

[15] S. M. Guo, L.C. Chen, J.S.H. Tsai. A boundary method for outlier detection based on support vector domain description[J]. Pattern Recognition, 2009, 42(1): 77-83.

[16] H. W. Cho. Data description and noise filtering based detection with its application and performance comparison[J]. Expert Systems with Applications, 2009, 36(1): 434-441.

[17] M. Arun Kumar, M. Gopal. Application of smoothing technique on twin support vector machines[J]. Pattern Recognition Letters, 2008, 29: 1842-1848.

[18] F. Sha, Y. Lin, L. K. Saul and D. D. Lee. Multiplicative updates for nonnegative quadratic programming[J]. Neural Computation, 2007, 19(8): 2004-2031.

[19] B. Schölkopf, A. Smola. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond[M]. Cambridge, MA: MIT Press, 2002.

[20] A. Asuncion, D. J. Newman. UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/ML Repository.html]. Irvine, CA: University of California, School of Information and Computer Science, 2007.