A new smooth method based on rotated hyperbola for support vector machine in classification

En Wang
School of Humanities and Social Science, Xian Jiaotong University, Shaanxi, 710049
E-mail: welg_glew@qq.com

Abstract. One smooth rotated hyperbola for support vector machine (RHSSVM) is proposed. Based on the approximation property of the hyperbola to its asymptotic lines, we build this model. Theoretical analysis shows that RHSSVM has the least error on approximating the plus function compared with three other typical models. Experimental results show that RHSSVM reduces the consuming time and efficiently deal with large scale and high dimensional programs.

1. Introduction
Based on Statistical Learning Theory (SLT), Support Vector Machine (SVM) performs its good proprieties in data classification and regression [1-3]. The SVM shows the performance of artificial neural networks in many areas such as text categorization, speech recognition, and scene classification, [4, 5]. But the objective function of SVM is non-differentiable at zero. In order to improve this situation, O. L. Mangasarian proposed an integral of the sigmoid function \( p(x,k) = x + \frac{1}{a} \log(1 + e^{-ax}) \) and got the smooth support vector machine (SSVM) [6]. In SSVM, we have the result of \( p(x,k)^2 - x^2 \leq 0.6927 \frac{1}{k^2} \), in which the plus function \( x = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases} \). Y. B. Yuan introduced two polynomial functions \( q(x,k) = \frac{k}{4} x^2 + \frac{1}{2} x + \frac{1}{4k} \) and \( h(x,k) = -\frac{k^3}{16} (x + \frac{1}{k})^3 (x - 3k) \), and got the quadratic polynomial smooth support vector machine (QPSSVM) and the fourth polynomial smooth support vector machine (FPSSVM) [7]. Then they drew the conclusions \( q(x,k)^2 - x^2 \leq 0.0909 \frac{1}{k^2} \), \( h(x,k)^2 - x^2 \leq 0.0526 \frac{1}{k^2} \). In 2005, Y. B. Yuan proposed the spline polynomial function and got the third-order spline support vector machine (TSSVM) [8]. Then the conclusion \( T(x,k)^2 - x^2 \leq 0.0417 \frac{1}{k^2} \) was come to.

In this paper, an unconstrained optimization model of smooth SVC is constructed firstly. Applying the smooth infinitely differentiable rotated hyperbola function \( h(x,k) \) to approximate plus function, we get the RHSSVM model. The global convergence of RHSSVM can also be proved. Because of the good properties, the fast Newton-Armijo algorithm [9] can be applied to train and test the smooth
RHSSVM. Meanwhile, we give the conclusion through theoretical analysis and experiments that RHSSVM is better than given models in smooth precision and classification capability.

2. The new RHSSVM model
Considering the programming of classifying \( m \) points in the \( n \)-dimensional real space \( R^n \), we have the corresponding label values \( y_i = \pm 1 \). The linear SVM optimization programming is given as follows

\[
\min_{w, b, \xi} H(w, b) = \frac{1}{2}(w^2 + b^2) + \nu \sum_{i=1}^{m} \xi_i
\]

subject to \( y_i(w \cdot x_i + b) + \xi_i \geq 1 \), \( \xi_i \geq 0, i = 1, 2, ..., m \),

where the \( w \) is weight vector, \( b \) is an offset and determines the distance to the origin point. The \( \nu > 0 \) it is a penalty parameter. In Formula (1), \( \xi_i \) is defined as \( \xi_i = \max\{0, 1 - y_i(w \cdot x_i + b)\} \), which is non-differentiable at zero. So we transfer the programming (1) to an unconstrained optimization problem

\[
\min_{w, b} H(w, b) = \frac{1}{2}(w^2 + b^2) + \nu \sum_{i=1}^{m} \max\{0, 1 - y_i(w \cdot x_i + b)\}.
\]

Let \( h(w, b) = \max\{0, 1 - y_i(w \cdot x_i + b)\} \). This function is similar to plus function. The plus function is continuous, convex but non-differentiable at the point \( x = 0 \). Therefore we cannot apply one fast algorithm to solve the programming (2). However, the simple rotated hyperbola function can be used to approximate the plus function when the two asymptotic lines have the intersection angle of \( 135^\circ \).

The hyperbola equation is \( y^2/a^2 - x^2/b^2 = 1 \ (y > 0) \), and the equations of its two asymptotic lines are \( y = \pm \frac{a}{b}x \). As the values of plus function are above zero, we can use the upper part of rotated hyperbola curve to approximate the plus function. Let \( \alpha = 22.5^\circ \), \( a = b \alpha \) in which \( p = \tan \alpha \), as the figure 1 shows.

\[\text{Figure 1. The simple hyperbola function.}\]

The rotated angle is \( \theta = 22.5^\circ \). So we can get the rotated hyperbola function

\[
y = \frac{1}{2}(a + \sqrt{b^2 + \frac{4a^2b^2}{b^2 \cos^2 \theta - a^2 \sin^2 \theta}}) = \frac{1}{2}(a + \sqrt{p^2 + \eta b^2}),
\]

where \( b = \beta / k \), (and \( \beta \) is a constant variable), \( \eta = \frac{4p^2}{\cos^2 \theta - p^2 \sin^2 \theta} \) ( \( \eta > 0 \) and \( \eta \) is also a constant variable), \( k \) is a smooth parameter. So we get the rotated hyperbola function
Replacing \( t \) with \((w, b)\), we will obtain the smooth function \( h(w, b; k) \)

\[
h(w, b; k) = \frac{1}{2} (1 - y_i(w \cdot x_i + b)) + \frac{1}{2} \sqrt{(1 - y_i(w \cdot x_i + b))^2 + \eta \beta^2}.
\]  

As we all know, the hyperbola function has the property of infinitely derivative. Thus, when we can utilize the rotated hyperbola function (3) to approximate plus function, the solution of the programming (2) can be solved as a smooth unconstrained optimization problem

\[
\min H(w, b; k) = \frac{1}{2} (w^2 + b^2) + \sum_{i=1}^{n} h(w, b; k).
\]

So far, we develop one new smooth rotated hyperbola support vector machine model (RHSSVM).

**Lemma** If the function \( H(w, b; k) \) is defined as (6), we have the following results.

(i) For any \( w \in \mathbb{R}^n \) and \( b \in \mathbb{R} \), \( H(w, b; k) \) is infinitely differentiable about \( w \) and \( b \).

(ii) For any \( w \in \mathbb{R}^n \), \( b \in \mathbb{R} \) and \( k^1, k^2 \) satisfy \( 0 < k^1 < k^2 \), we have \( H(w, b; k^1) > H(w, b; k^2) \).

(iii) For any \( k > 0 \), \( H(w, b; k) \) is continuously differentiable and strictly convex.

(iv) For any \( k > 0 \), \( h(w, b; k) = \frac{1}{2} (w^2 + b^2) + \sum_{i=1}^{n} h(w, b; k) \) at the range of \( \left[ -\frac{1}{k}, \frac{1}{k} \right] \), we have

\[
h(w, b; k) = \frac{1}{2} (w^2 + b^2) + \sum_{i=1}^{n} h(w, b; k).
\]

Proof. (i) It is easy to obtain that the first and second order derivatives of \( h(t, k) \):

\[
\frac{\partial h(x, k)}{\partial x} = \frac{1}{2} (1 + x(x^2 + \eta \beta^2)^{-\frac{1}{2}}), \quad (-\infty < x < +\infty)
\]  

and

\[
\frac{\partial^2 h(x, k)}{\partial x^2} = \frac{1}{2} x(x^2 + \eta \beta^2)^{-\frac{3}{2}}, \quad (-\infty < x < +\infty).
\]

From (7) and (8), we can gain that \( \frac{d^2 h(x, k)}{dx^2} \) exists. Therefore, according to (4) and (5), the result that \( H(w, b; k) \) is infinitely differentiable about \( w \) and \( b \) holds.

(ii) For any \( t \in \mathbb{R}^n \) and \( k > 0 \), we have

\[
\frac{\partial h(t; k)}{\partial k} = \frac{\eta \beta^2}{2k^3 (\eta \beta^2 + \frac{t^2}{k^2})} < 0, \quad \frac{\partial h(t; k)}{\partial t} = \frac{1}{2} \left( \frac{\eta \beta^2}{\sqrt{\eta \beta^2 + \frac{t^2}{k^2}}} \right) > 0.
\]

From above, we know that the function \( h(t; k) \) is monotone about \( k \). For any \( 0 < k^1 < k^2 \), we have the inequality \( h(t; k^1) > h(t; k^2) \). From the definition of \( H(w, b; k) \), the conclusion holds obviously.

(iii) Now we prove the function \( H(w, b; k) \) is strictly convex. For any \( k > 0 \), \( H(w, b; k) \) is continuously differentiable. From (5) and (6), we know that

\[
\nabla_x H(w, b; k) = -\frac{\eta \beta^2}{2k^3 (\eta \beta^2 + \frac{t^2}{k^2})} \frac{1 - y_i(w \cdot x_i + b)}{(1 - y_i(w \cdot x_i + b))^2 + \eta \beta^2} x_i,
\]

\[
\nabla_t H(w, b; k) = -\frac{\eta \beta^2}{2k^3 (\eta \beta^2 + \frac{t^2}{k^2})} \frac{1 - y_i(w \cdot x_i + b)}{(1 - y_i(w \cdot x_i + b))^2 + \eta \beta^2} x_i.
\]

It is easy to get
\[ V^2 H(w,b;k) = \begin{cases} I_x + \sum_{i=1}^{n} \lambda_i(w,b;k)x_i^2 & \sum_{i=1}^{m} \lambda_i(w,b;k)x_i \\ \sum_{i=1}^{m} \lambda_i(w,b;k)x_i^2 & \sum_{i=1}^{m} \lambda_i(w,b;k) \end{cases}, \]  
where \( \lambda_i(w,b;k) = \frac{\eta \nu \beta^2}{2k^2 \left( (1 - y_i(w \cdot x_i + b))^2 + \eta \beta^2 \right)^\frac{1}{2}} \), and \( x_i^\top \) represents the transpose of the \( x \). For any \( \bar{x} \in R^m \) and \( \bar{z} \neq 0, \bar{z} \in R^n \), we have
\[
\bar{z}^\top V^2 H(w,b;k) \bar{z} = \bar{z}^\top \bar{z} + \sum_{i=1}^{m} \lambda_i(w,b;k)(x_i^\top \bar{z} + \bar{z}^\top x_i)^2 + 2 \sum_{i=1}^{m} \lambda_i(w,b;k) \bar{z}^\top x_i \bar{z} + \sum_{i=1}^{m} \lambda_i(w,b;k) \bar{z}^\top \bar{z}
\]
where \( \lambda_i(w,b;k) > 0 \). It shows that \( H(w,b;k) \) is a strictly convex function for any \( k > 0 \).

(iv) For \( x \in [-\frac{1}{k}, 0], \)
\[ h^2(w,b;k) - x_i^2 = h(x,k)^2. \]  
As \( h(x,k)^2 > 0 \), we have \( h(x,k)^2 - x_i^2 = h(x,k)^2 \leq h(0,k)^2 = \frac{\eta \nu \beta^2}{4k^2} \). For \( x \in [0, \frac{1}{k}], \)
\[ h^2(x,k) - x_i^2 \leq \frac{\eta \nu \beta^2}{4k^2} + \frac{1 + \eta \nu \beta^2}{2k^2} - 1 \]  
holds. From above, we have
\[ h^2(x,k) - x_i^2 \leq \frac{\eta \nu \beta^2}{4k^2} + \frac{1 + \eta \nu \beta^2}{2k^2} - 1 = 0.0103 \leq \frac{1}{k^2}. \]

So far, the lemma holds.

From the figure 2, we can see that RHSSVM has the best approximation performance to plus function.

![Figure 2. Smooth performance comparisons with the RHSSVM.](image)

3. Convergence performance of RHSSVM
In this part, we present the convergence of RHSSVM. We will prove that the solution of RHSSVM can closely approximate the optimal solution of the original model (2) when \( k \) draws close towards positive infinity. In addition, the convergence condition holds in the nonlinear RHSSVM.

**Theorem** Let \( A \in R^{m \times n}, b \in R^n \), define real functions \( t(x) \) and \( f(x,k) \) as follows,
\[ t(x) = \frac{1}{2} \| Ax - b \|_2^2 + \frac{1}{2} \| x \|_2^2, \]
\[ f(x,k) = \frac{1}{2} \| h(Ax - b, k) \|_2^2 + \frac{1}{2} \| x \|_2^2. \]  

Then next results can be achieved,
(i) \( t(x) \) and \( f(x,k) \) are strongly convex functions;
(ii) There is a unique solution \( x^* \) to \( \min_{x \in \mathbb{R}^n} t(x) \) and a unique solution \( x_i^* \) to \( \min_{x \in \mathbb{R}^n} f(x, k) \);

(iii) \( \forall k > 0 \), we have \( \|x_i^* - x^*\| \leq \frac{0.00515m}{k^2} \) when the variable \( \beta = 0.1 \).

(iv) \( \lim_{k \to \infty} \|x_i^* - x^*\| = 0 \).

Proof. (i) For arbitrary \( k > 0 \), \( t(x) \) and \( f(x, k) \) are strongly convex functions because \( \|\| \|^2 \) is a strong convex function.

(ii) Let \( L_{t}(t(x)) \) be the level set of \( t(x) \) and let \( L_{f}(f(x, k)) \) be the level set of \( f(x, k) \). As \( f(x, k) \geq x^* \), it is easy to obtain \( L_{f}(f(x, k)) \subseteq L_{t}(t(x)) \subseteq \{x \mid \|x\|^2 \leq 2\nu \} \). Therefore, \( L_{t}(t(x)) \) and \( L_{f}(f(x, k)) \) are compact subsets in \( \mathbb{R}^n \). When we apply the strong convexity property of \( t(x) \) and \( f(x, k) \), there is a unique solution for both of \( \min_{x \in \mathbb{R}^n} t(x) \) and \( \min_{x \in \mathbb{R}^n} f(x, k) \).

(iii) Using the first order optimization condition and considering the convex property of \( t(x) \) and \( f(x, k) \), we can attain

\[
t(x_i^*) - t(x^*) \geq \nabla t(x^*)(x_i^* - x^*) + \frac{1}{2} \|x_i^* - x^*\|^2 = \frac{1}{2} \|x_i^* - x^*\|^2.
\]

\[
f(x^*, k) - f(x_i^*, k) \geq \nabla f(x^*)(x_i^* - x^*) + \frac{1}{2} \|x_i^* - x^*\|^2 = \frac{1}{2} \|x_i^* - x^*\|^2.
\]

When adding the two formulas above and paying attention to \( f(x, k) \geq x^* \), we can get

\[
\|x_i^* - x^*\| \leq t(x_i^*) - t(x^*) + f(x^*, k) - f(x_i^*, k)
\]

\[
= f(x^*, k) - t(x^*) - f(x_i^*, k) - t(x_i^*)
\]

\[
\leq f(x^*, k) - t(x^*)
\]

\[
= \frac{1}{2} \|h(Ax - b, k)\|^2 - \frac{1}{2} \|h(Ax - b, k)\|^2.
\]

According to the second result of Lemma 2, we know that \( \|x_i^* - x^*\| \leq \frac{0.00515m}{k^2} \).

(iv) As \( \|x_i^* - x^*\| \leq \frac{0.00515m}{k^2} \), we can get \( \lim_{k \to \infty} \|x_i^* - x^*\| = 0 \) easily.

4. Experiment results and comparisons

This section presents the comparisons among SSVM, FPSSVM, TSSVM, and RHSSVM. The SSVM applies the sigmoid function to replace the plus function. The FPSSVM and TSSVM utilize the fourth polynomial function and spline function to approximate the plus function. For the four models, we utilize the tolerance \( 10^{-3} \) to terminate the program execution and we perform tenfold cross-validation on each dataset. For the nonlinear data, we apply the Gaussian kernel.

The first experiment is simulated on the UCI linear datasets [10] and the original SVM codes are from Lee and his team.

Table 1 shows that RHSSVM has the best classification property than the other threes models. Bold type indicates the best results.

The second experiment is based on the non-linear datasets the “tried and true” checkerboard dataset [11]. The fourth experiment is the checkboard dataset which is one nonlinearly separable dataset. From the two-dimensional direction, we sample 100 points at the range of 0–99 randomly. Then, we get the 10,000 points and its corresponding two classes “White” and “Black” spaced by 4×4 grids.

In the first trial of this experiment, the training set contains 1000 points randomly sampled from the checkerboard which contains 514 “white” samples and 486 “black” samples. The rest 9,000 points are
in the testing set. At the same time, we utilize the Gaussian kernel function $K(x, y) = \exp(-\mu \|x - y\|)$ to map the data into high dimensions. Table 2 shows that RHSSVM has much less computing time and the accuracy is better than the SSVM, FPSSVM, and TSSVM.

$$K(x, y) = \exp(-\mu \|x - y\|)$$

| Datasets            | Algorithm | Time (CPU sec) | Train Correctness | Test Correctness |
|---------------------|-----------|----------------|-------------------|------------------|
| Bupa                | SSVM      | 0.02           | 70.11%            | 65.78%           |
|                     | FPSSVM    | 0.02           | 70.08%            | 65.78%           |
|                     | TPSSVM    | 0.02           | 70.08%            | 65.78%           |
|                     | RHSSVM    | 0.02           | 70.76%            | 66.36%           |
| Pima diabetes       | SSVM      | 0.03           | 77.97%            | 77.72%           |
|                     | FPSSVM    | 0.02           | 77.95%            | 77.72%           |
|                     | TPSSVM    | 0.03           | 77.95%            | 77.59%           |
|                     | RHSSVM    | 0.03           | 77.95%            | 77.73%           |
| Ionosphere          | SSVM      | 0.05           | 93.86%            | 86.58%           |
|                     | FPSSVM    | 0.04           | 93.86%            | 86.58%           |
|                     | TPSSVM    | 0.04           | 93.86%            | 86.58%           |
|                     | RHSSVM    | 0.07           | 93.86%            | 87.15%           |
| WDBC                | SSVM      | 0.05           | 98.93%            | 96.48%           |
|                     | FPSSVM    | 0.05           | 98.93%            | 96.48%           |
|                     | TPSSVM    | 0.06           | 98.93%            | 96.48%           |
|                     | RHSSVM    | 0.08           | 99.10%            | 97.01%           |
| Adult               | SSVM      | 0.78           | 84.12%            | 84.10%           |
|                     | FPSSVM    | 0.75           | 84.12%            | 84.09%           |
|                     | TPSSVM    | 0.83           | 84.12%            | 84.09%           |
|                     | RHSSVM    | 0.72           | 84.24%            | 84.16%           |

The results in table 2 demonstrate that RHSSVM can solve massive problems with the highest efficiency, followed by TSSVM, FPSVM, and SSVM. For the magnitude datasets, small improvement on classification can decrease quantities of wrong judgment in our real world. The experimental results show that RHSSVM can obtain the least computing time and best test precision.
5. Conclusions
Utilizing the good propriety of the rotated hyperbola, we get the RHSSVM. The new model shows its good classification performance, less computing time, and best expansibility to massively sized datasets on both artificial datasets and real-world datasets. For the nonlinear RHSSVM model, with the growth of the size of the training dataset, the consuming time decreases sharply. Therefore, the proposed RHSSVM is more suitable to deal with the nonlinear and real-world datasets.

Acknowledgments
This work was supported in part by the Nature Science Foundation of China under Grant (61100165, 61100231, 61472307) and Natural Science Foundation of Shaanxi Province (2016JM6004).

References
[1] Burges C J C 1998 A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* 2:121-167
[2] Huang HJ, Ding SF and Shi ZZ 2015 Smooth CHKS twin support vector regression. *Journal of Computer Research & Development* 52:561-568
[3] J Ind Manag 2016 Fuzzy quadratic surface support vector machine based on fisher discriminant analysis. *Optimization* 12: 357-373
[4] Li K L, Xie J and Sun X 2011 Multi-class text categorization based on LDA and SVM. *Procedia Engineering* 15: 1963-1967.
[5] Chandaka Suryanarayana 2009 Support vector machines employing cross-correlation for emotional speech recognition. *Measurement* 42(4):611-618.
[6] Lee Y J and Mangarasian O L 2001 SSVM. A smooth support vector machine for classification. *Computational Optimization and Applications* 20: 5-22
[7] Yuan Y B, Yan J and Xu C 2005 Polynomial smooth support vector machine: PSSVM. *Chinese Journal of Computers* 28: 9-17
[8] Yuan Y B, Fan W and Pu D 2007 Spline function smooth support vector machine for classification. *Journal of Industrial and Management Optimization* 3: 529-542
[9] Bertsekas D P 1999 Nonlinear programming. Athena Scientific. Belmont. MA. second edition
[10] UC Irvine Machine Learning Repository. http://archive.ics.uci.edu/ml/datasets/Liver Disorders
[11] Ho T K and Kleinberg E M 1996 “Checkerboard dataset” ftp://ftp.cs.wisc.edu/math-prog/cpo-dataset/machine-learn/checker