A New Smooth Support Vector Machine with 1-Norm Penalty Term

http://dx.doi.org/10.3991/ijoe.v9iS4.2657

J.D. Shen[1], X.J. Peng[2]
[1] China Jiliang University, Hangzhou, P R China
[2] Shanghai Normal University, Shanghai, P R China

***Abstract*—Recently, soft margin smooth support vector machine with 1-norm penalty term ($SSVM_1$) is discovered to possess better outlier resistance than soft margin smooth support vector machine with 2-norm penalty term ($SSVM_2$). One of the most important steps in the framework of SSVMs is to replace the x_+ by a differential function in the primal model, and get an approximate solution. This study proposes one function constructed by Padé approximant via the formal orthogonal polynomials as the smoothing technique, and a new 1-norm SSVM, Padé $SSVM_1$, is represented. A method for outlier filtering is proposed to improve the ability of outlier resistance. The experimental results show that Padé $SSVM_1$, even without outlier filtering, performs better than the previous $SSVM_2$ and $SSVM_1$ on the polluted synthetic datasets.**

***Index Terms*—Smooth support vector machine, Padé approximant, Outlier resistance, 1-norm**

I. INTRODUCTION

Support vector machines (SVMs) have been proven to be one of the promising learning algorithms for classification [1]. The standard SVMs have loss + penalty terms measured by 1-norm or 2-norm measurements. The loss part measures the quality of model fitting and the penalty part controls the model complexity. In [2], Li-Jen Chien et al. showed that the measurement of the 2-norm loss term amplifies the effect of outliers much more than the measurement of the 1-norm loss term in training process. From this robustness point of view, the authors in [2] developed a $SSVM_1$ whose loss term is measured by 1-norm and the integral of the sigmoid function was selected as the smoothing technique (Sigmoid $SSVM_1$ for short). Finally, the experiments in [2] showed that Sigmoid $SSVM_1$ can remedy the drawback of 2-norm soft margin smooth support vector machine ($SSVM_2$) [3] for outlier effect and thus get outlier resistance.

Although SVMs have the advantage of being robust for outlier effect [4], there are still some violent cases that will mislead SVM classifiers to lose their generalization ability for prediction, even the good sigmoid $SSVM_1$ also became powerless at this time. Li-Jen Chien, Y.J. Lee, Z. P. Kao, and C. C. Chang [2] proposed a heuristic method to filter outliers among Newton-Armijo iteration of the training process and make SSVMs be more robust while encountering datasets with extreme outliers.

In this study, we will give a new smoothing technique, Padé approximant, which can approximate the plus function $x_+ = \max\{x; 0\}$ more accurately than the integral of the sigmoid function. The $SSVM_1$ smoothed by this function is denoted by Padé $SSVM_1$. We will show that the outlier resistance of Padé $SSVM_1$ is better than that of Sigmoid $SSVM_1$ in most of the cases, even still performs well in those violent cases. We will also give another strategy for outlier filtering, which turns out to be efficient to make $SSVM_2$ and Sigmoid $SSVM_1$ be robust for those datasets polluted with extreme outliers.

II. 1-NORM SOFT SVM ($SSVM_1$)

Consider the binary problem of classifying m points in the n-dimensional real space R^n, represented by an $m \times n$ matrix A. According to membership of each point $A_i \in R^{n\times 1}$ in the classes +1 or -1, D is an $m \times m$ diagonal matrix with ones or minus ones along its diagonal. Similar to the framework of $SSVM_2$ [3], the classification problem can be reformulated as follows:

$$\begin{aligned}&\min_{(w,b,\xi)\in R^{(n+1+m)}} \frac{1}{2}(\|w\|_2^2 + b^2) + C\|\xi\|_1\\ &\text{subject to: } D(Aw + 1b) + \xi \geq 1\,,\\ &\qquad\qquad \xi \geq 0\end{aligned} \tag{1}$$

As a solution of problem (1), the slack variable ξ is given by

$$\xi = (1 - D(Aw + 1b))_+, \tag{2}$$

Thus, we can replace ξ in constraint (1) by (2) and convert the SVM problem (1) into an equivalent SVM which is an unconstrained optimization problem as follows:

$$\min_{(w,b)\in R^{(n+1)}} \frac{1}{2}(\|w\|_2^2 + b^2) + C\|(\mathbf{1} - D(Aw + \mathbf{1}b))_+\|_1. \tag{3}$$

The problem is a strongly convex minimization problem without any constraint. Thus, problem (3) has a unique solution. Obviously, the objective function in (3) is not twice differentiable which precludes the use of a fast Newton method, because it always requires the objective function′s gradient and Hessian matrix. Y. J. Lee and O. L. Mangasarian [3] applied the smoothing technique and replaced x_+ by the integral of the sigmoid function $1/(1+e^{-\eta x})$ of neural networks:

$$\rho(x,\eta) = x + \frac{1}{\eta}\ln(1+e^{-\eta x}), \quad \eta > 0. \tag{4}$$

This ρ function with a smoothing parameter η is used here to simultaneously smooth and approximate the model (3), i.e., we use a differential (twice differentiable at least) function ρ to replace the plus function $(\cdot)_+$ in (3) in order to get an approximate solution of the model. Finally, we obtain the 1-norm smooth support vector machine with respect to the integral of the sigmoid function (Sigmoid $SSVM_1$ for short):

$$\min_{(w,b)\in R^{(n+1)}} \frac{1}{2}(\|w\|_2^2 + b^2) + C\|\rho(\mathbf{1} - D(Aw+\mathbf{1}b),\eta)_+\|_1 . \tag{5}$$

By taking the advantage of the twice differentiability of the objective functions on problem (5), a prescribed quadratically convergent Newton-Armijo algorithm [5] can be used to solve this problem. Hence, the smoothing problem can be solved without a sophisticated optimization solver.

The transformation from (3) to (5) raises a very natural question: Are the two models equivalent? In fact, the model after smoothing is not equal to the primal problem (3) anymore. But in an analogous manner as in [3], it is easy to be proved that the solution of (5) converges to the unique solution of the primal problem when the smoothing parameter η in the $SSVM_1$ approaches infinity. It is just because of the truth: if the value of η increases, the $\rho(x,\eta)$ will approximate the plus function more accurately. Therefore, how to construct an efficient smoothing technique to achieve the simultaneous smoothing and approximation naturally becomes the major goal of this study.

III. 1-NORM SMOOTH SUPPORT VECTOR MACHINE BASED ON PADÉ APPROXIMANT

In this section, we propose a kind of rational function, namely Padé approximant, as the smoothing technique to simultaneously smooth and approximate the plus function in the framework of $SSVM_1$.

A. Padé Approximation via the FOP

Let $f(x)$ be a given power series with coefficients $c_i \in C$,

$$f(x) = c_0 + c_1 x + c_2 x^2 + \cdots + c_n x^n + \cdots, \tag{6}$$

For above $f(x)$, we give the definition of Padé approximation as follows.

Definition 3.1. Let $\tilde{Q}_n(x)$ and $\tilde{P}_m(x)$ be two polynomials of degree m and n respectively, if the following relation holds:

$$\tilde{Q}_n(x) f(x) - \tilde{P}_m(x) = O\left(x^{m+n+1}\right), \tag{7}$$

where the right-hand side denotes a power series in x with lowest order term of degree m+n+1 or higher, then $\tilde{P}_m(x)/\tilde{Q}_n(x)$ is called Padé approximant for $f(x)$ and is denoted by $[m/n]f(x)$.

Let $c^{(h)}$: P→C be a linear functional on the polynomial space P, which is defined by

$$c^{(h)}\left(t^i\right) = c_{h+i}, \quad i = 0,1,\cdots, \tag{8}$$

where

$$c^{(0)}\left(t^i\right) \triangleq c\left(t^i\right) = c_i, \quad i = 0,1,\cdots, \tag{9}$$

with the convention that $c_i = 0$ for $i < 0$.

We now give the definition of formal orthogonal polynomials (FOPs) associated with $c^{(m-n+1)}$, which is defined by [7] with h=m-n+1.

Definition 3.2. $\{q_k\}$ is called a family of formal orthogonal polynomials associated with $c^{(m-n+1)}$ if, $k \ge 0$, q_k has degree k at most and

$$c^{(m-n+1)}\left(t^i q_k(t)\right) = 0, \quad i = 0,\cdots,k-1, \tag{10}$$

Now we present a main theorem (its proof is referred to [8]) about Padé approximation via the formal orthogonal polynomials (PAVOP) as follows.

Theorem 3.3. Let q_n be a polynomial which belongs to the family of formal orthogonal polynomials associated with $f(x)$,

$$q_n(t) = a_0 + a_1 t + \cdots a_n t^n = \sum_{i=0}^{n} a_i t^i, \tag{11}$$

satisfies

$$c^{(m-n+1)}\left(t^i q_n(t)\right) = 0, \quad i = 0,\cdots,n-1, \tag{12}$$

and set

$$\tilde{Q}_n(t) = t^n q_n\left(t^{-1}\right) = \sum_{i=0}^{n} a_i t^{n-i}.. \tag{13}$$

Define the polynomial $\tilde{P}_m(x)$

$$\tilde{P}_m(m) = \sum_{i=0}^{n} a_i x^{n-i} f_{m-n+i}(x), \tag{14}$$

where

$$f_k(x) = \begin{cases} \sum_{j=0}^{k} c_j x^j, & k \ge 0 \\ 0, & k < 0 \end{cases} \tag{15}$$

Then, it holds

$$[m/n]_f(x)=\frac{P_m(x)}{Q_n(x)}=\frac{\sum_{i=0}^{n}a_i x^{n-i}f_{m-n+i}(x)}{\sum_{i=0}^{n}a_i x^{n-i}}, \tag{16}$$

That is,

$$Q_n(x)f(x)-P_m(x)=O\left(x^{m+n+1}\right). \tag{17}$$

B. *Padé Approximant for* x_+

We now consider using a Padé approximant to simultaneously smooth and approximate the plus function x_+.

It is well known that the plus function is not smooth, but continuous, so we can expand the plus function to a power series:

$$\begin{aligned} x_+ &= \frac{|x|+x}{2} \\ &= \frac{1}{2\eta}\left[\frac{1+\eta^2x^2}{2}-\sum_{n=2}^{\infty}\frac{(2n-3)!!}{(2n)!!}\left(1-\eta^2x^2\right)^n\right]+\frac{x}{2}. \end{aligned} \tag{18}$$

Then a Padé approximant for the above power series is computed by Thereom 3.3:

$$\frac{1}{2\eta}g\frac{1+10\eta^2x^2+5\eta^4x^4}{5+11\eta^2x^2+\eta^4x^4}+\frac{x}{2}. \tag{19}$$

Now we first give the smooth function whose main component is just the Padé approximant (19):

$$P(x,\eta)=\begin{cases} x, & x\ge\frac{1}{\eta} \\ \frac{1}{2\eta}g\frac{1+10\eta^2x^2+5\eta^4x^4}{5+11\eta^2x^2+\eta^4x^4}+\frac{x}{2}, & -\frac{1}{\eta}<x<\frac{1}{\eta} \\ 0, & x\le-\frac{1}{\eta} \end{cases} \tag{20}$$

and then a Padé $SSVM_1$ model is constructed:

$$\min_{(w,b)\in R^{(n+1)}}\frac{1}{2}(\|w\|_2^2+b^2)+C\|P(\mathbf{1}-D(Aw+\mathbf{1}b),\eta)\|_1, \tag{21}$$

where **1** denotes a column vector of ones for arbitrary dimension, and function P has an effect on all components of a matrix or a vector in (21), i.e., $P(\mathbf{1}-D(Aw+\mathbf{1}b),\eta)\in R^m$, $(P(\mathbf{1}-D(Aw+\mathbf{1}b),\eta))_i=P(1-D_i(A_iw+b),\eta)$, and η whose value is not a main factor for the final $SSVM_1$ is called smoothing parameter. We will now show a simple theorem that bounds the difference between the plus function x_+ and its smooth approximant $P(x,\eta)$.

Theorem 3.4. Let $x\in R$, $P(x,\eta)$ are defined as (20), x_+ is the plus function:

(i) $P(x,\eta)$ is quadratic smoothness, at the point $x=\pm1/\eta$, $x=0$, satisfies:

$$\begin{cases} P(\frac{1}{\eta},\eta)=\frac{1}{\eta}, P(-\frac{1}{\eta},\eta)=0; \\ \nabla P(\frac{1}{\eta},\eta)=1, \nabla P(-\frac{1}{\eta},\eta)=0; \\ \nabla^2 P(\frac{1}{\eta},\eta)=1, \nabla^2 P(-\frac{1}{\eta},\eta)=0; \end{cases} \tag{22}$$

(ii)

$$P(x,\eta)>x_+; \tag{23}$$

(iii) for arbitrary x, η

$$P(x,\eta)-x_+\le 0.100/\eta. \tag{24}$$

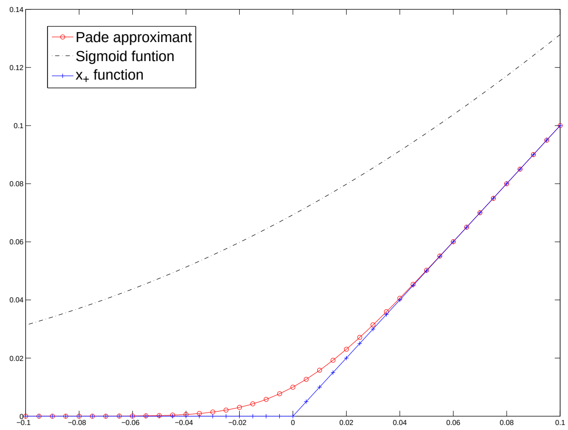

Figure 1. The approximation of two smooth functions to x_+, with η= 10.

The Newton-Armijo algorithm with respect to $SSVM_1$ is omitted here because it is running the same procedure as that in 2-norm problem.

IV. NUMERICAL RESULTS AND A METHOD FOR OUTLIER FILTERING

As stated in [9], Sigmoid $SSVM_1$ possesses good outlier resistance, which can be observed in a numerical tests. The first result is represented in Fig. 2 and the corresponding comparison of correctness is in Table I.

As has been already pointed out by Li-Jen Chien, there are some violent cases that are still easy to mislead either Sigmoid $SSVM_1$ or Sigmoid $SSVM_2$ to lose their generalization ability. A violent case is presented in Fig. 3, similar with Fig. 1 in [2], in which the positive and negative are normal distribution with mean 2 and -2 respectively and deviation 1. The outlier difference is 75 from the mean and the outlier ratio is 0.025 in positive and negative totally. In this case, no matter Sigmoid $SSVM_2$ or Sigmoid $SSVM_1$, both of them lost efficacy. Why all of the SVMs (Sigmoid $SSVM_1$, Sigmoid $SSVM_2$, including LIBSVM [10]) lose their generalization ability in this case is that they pay too much effort to minimize the loss term and sacrifice for minimizing the penalty term because of these extreme outliers [2]. Fortunately, Padé $SSVM_1$ is still robust, and attains the generalization in this violent case.

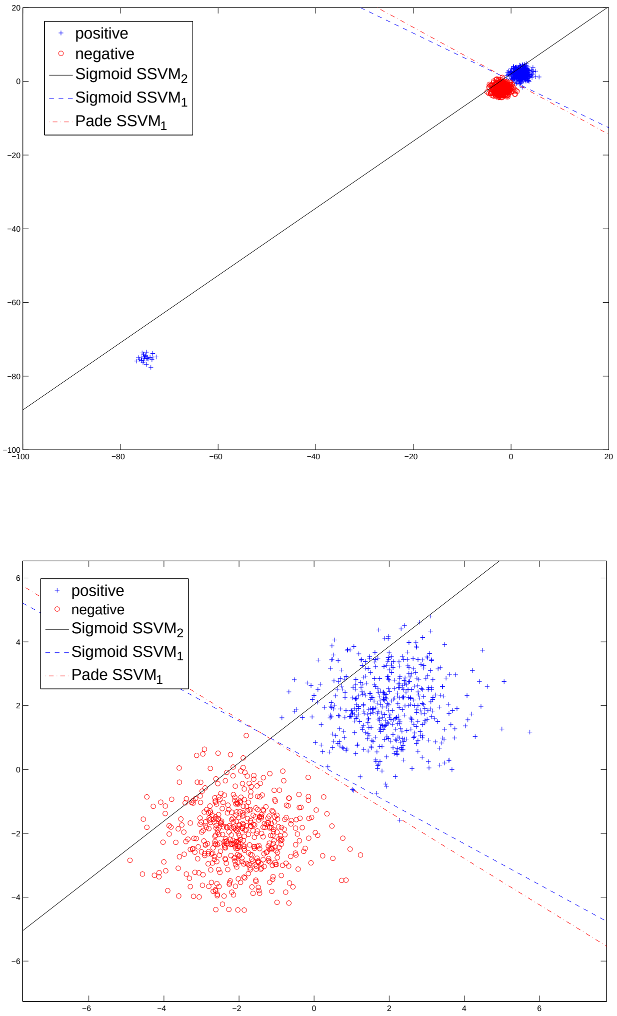

Figure 2. The synthetic dataset: a normal distribution, mean = 2 and -2, the standard deviation = 1. The outlier ratio is 0.025 in the positive examples, and outliers are on the lower-left corners in the panorama. For the outliers, the outlier difference from the mean of positive groups is set to be 75 times the standard deviation.

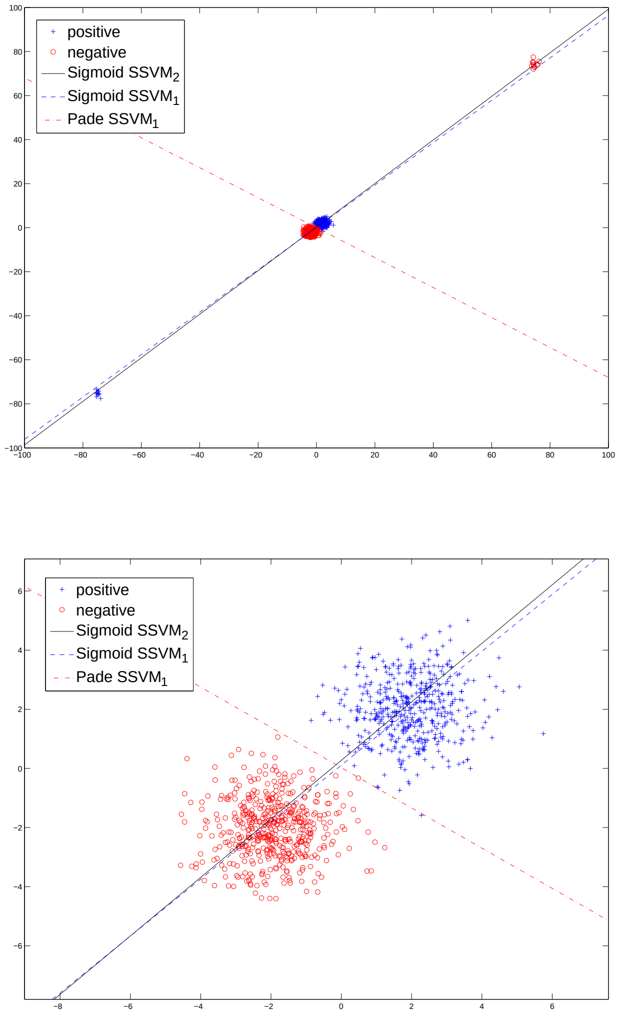

Figure 3. The positive and negative is the same normal distribution as in Fig. 2. The outlier ratio is 0.025 in positive and negative examples, and outliers are on the upper-right and lower-left corners in the left figure). For the outliers, the outlier difference from the mean of their groups is set to be 75 times the standard deviation.

TABLE I.
CORRECTNESS OF THREE SSVMS IN ABOVE EXPERIMENT

Method	10-fold training correctness, %	10-fold testing correctness, %
Sigmoid SSVM$_2$	51.7140	48.4000
Sigmoid SSVM$_1$	80.4627	78.8000
Padé SSVM$_1$	97.5184	95.2000

TABLE II.
CORRECTNESS OF THREE SSVMS IN ABOVE EXPERIMENT

Method	10-fold training correctness, %	10-fold testing correctness, %
Sigmoid SSVM$_2$	52.6526	55.6000
Sigmoid SSVM$_1$	53.7684	52.8000
Padé SSVM$_1$	97.2632	97.2000

To eliminate the influence of outliers in such violent case, Li-Jen Chien, Y.J. Lee, Z. P. Kao, and C. C. Chang [2] prescribed a heuristic method to filter out the extreme outliers. In this study, we give another slightly different strategy to filter out the extreme outliers. We would first run the process of SSVM$_1$, and then ignore some large ξ_i's. But how to determine the value of ξ_i is large enough? We set outlier ratio as our threshold. In our method, the samples whose ξ_i's are over 90 percentage are ignored until the threshold reaches the outlier ratio, and finally we use the rest samples to reconstruct a new SSVM$_1$ as the final classifier. We denote this outlier filtering method by SSVM$_{1-o}$.

Fig. 4 are in the same setting as Fig. 3. It is very obvious that SSVM$_{1-o}$ and SSVM$_{2-o}$ successfully classify the most of examples. But among them, Padé SSVM$_{1-o}$ performs the best.

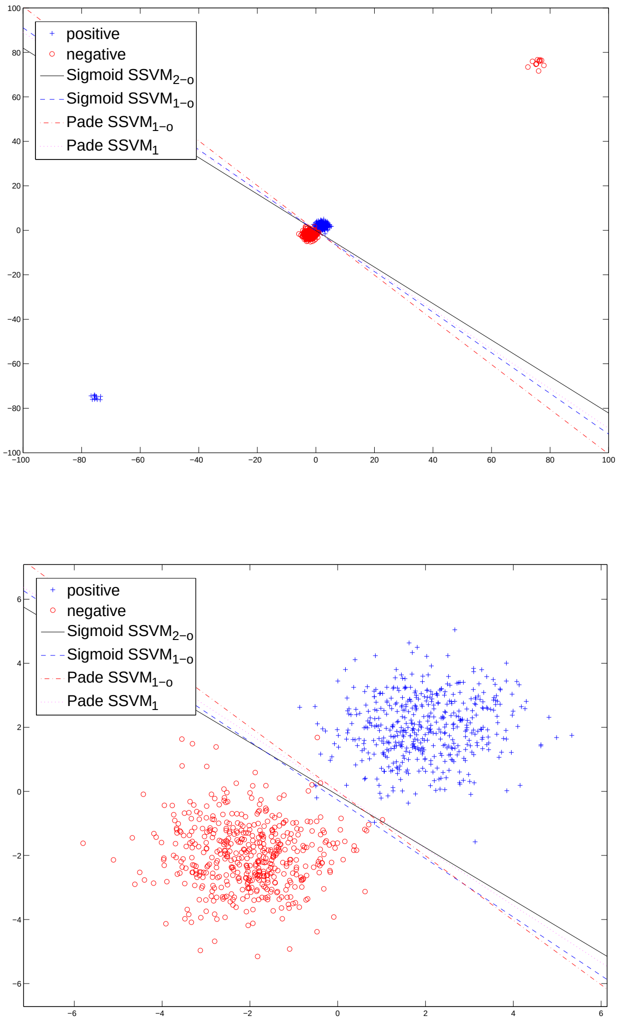

Figure 4. The same violent case classified by SSVM$_{2-o}$, Sigmoid SSVM$_{1-o}$ and Padé SSVM$_{1-o}$

V. CONCLUSIONS

We have proposed Padé approximant as a new smoothing technique for SSVM$_1$. The new SSVM$_1$ constructed by this Padé approximant, i.e., Padé SSVM$_1$, has been proved by the theoretical analyses and the numerical results to possess the best outlier resistance compared with previous SSVMs. To strengthen the robustness of SSVMs in some violent cases, a simple method for outlier filtering is proposed. This method for outlier filtering also improves robustness a lot for Sigmoid SSVM$_1$ and Sigmoid SSVM$_2$.

ACKNOWLEDGEMENT

The research was supported by the National Natural Science Foundation of China (No.61003207, 61202156), the Natural Science Foundation of Zhejiang Province of China (No.Y6100588), and the Natural Science Foundation of Shanghai (12ZR1447100).

REFERENCES

[1] V. N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons , USA, 1998.

[2] L. J. Chein, Y.J. Lee, Z. P. Kao, and C. C. Chang, "Robust 1-norm soft margin smooth support vector machine", In: C. Fyfe et al. (Eds.): IDEAL 2010, LNCS 6283, Springer-Verlag Berlin Heidelberg, 2010, pp. 145-152.

[3] Y. J. Lee and O. L. Mangasarian, "SSVM: A smooth support vector machine", Comput. Optim. Appl., vol. 20, pp. 5-22, 2001. http://dx.doi.org/10.1023/A:1011215321374

[4] H. Xu, C. Caramanis and S. Mannor, "Robustness and regularization of support vector machines", J. Mach. Learn. Res., vol. 10, pp. 1485-1510, 2009.

[5] D. P. Bertsekas, *Nonlinear programming*, 2nd edt., Athena Scientific, Belmont, 1999.

[6] O. L. Mangasarian, "Mathematical programming in neural networks", ORSA Journal on Computing, vol. 5, no. 4, pp. 349-360, 1993. http://dx.doi.org/10.1287/ijoc.5.4.349

[7] C. Brezinski, *Computational aspects of linear control*, Kluwer Academic Publishers, The Netherlands, 2002. http://dx.doi.org/10.1007/978-1-4613-0261-2

[8] C. Q. Gu and J. D. Shen, "Function-valued Padé-type approximant via the formal orthogonal polynomials and its applications in solving integral equations", J. Comput. Appl. Math., vol. 221, no. 1, pp. 114-131, 2008. http://dx.doi.org/10.1016/j.cam.2007.10.008

[9] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, 2000.

[10] LIBSVM, a library for support vector machines (http://www.csie.ntu.edu.tw/ cjlin/libsvm).

AUTHORS

J. D. Shen is with the Department of Information and Mathematics Sciences, China Jiliang University Department of Information and Mathematics Sciences, China Jiliang University, Hangzhou 310018, Zhejiang Province, P R China. (e-mail: cs.jindong@ gmail.com).

X. J. Peng is with the Department of Mathematics, Shanghai Normal University, Shanghai, 200234, P R China (e-mail: xjpeng999@gmail.com).

This work was supported by the National Natural Science Foundation of China (No.61003207, 61202156), the Natural Science Foundation of Zhejiang Province of China (No.Y6100588), and the Natural Science Foundation of Shanghai (12ZR1447100). This article is an extended and modified version of a paper presented at the International Conference on Mechanical Engineering, Automation and Material Science (MEAMS2012), held 22-23 December 2012, Wuhan, China. Received 09 January 2013. Published as resubmitted by the authors 01 May 2013.