Intelligent Information Management, 2012, 4, 45-51
http://dx.doi.org/10.4236/iim.2012.43007 Published Online May 2012 (http://www.SciRP.org/journal/iim)

The Asymptotic Study of Smooth Entropy Support Vector Regression

Guo-Sheng Hu, Jian-Hua Zhang
Shanghai Technical Institute of Electronics and
Information, School of Electronics and Information, Shanghai, China
Email: jamhu8@sohu.com

Received December 4, 2011; revised February 22, 2012; accepted February 28, 2012

ABSTRACT

In this paper, a novel formulation, smooth entropy support vector regression (SESVR), is proposed, which is a smooth unconstrained optimization reformulation of the traditional linear programming associated with a ε-insensitive support vector regression. An entropy penalty function is substituted for the plus function to make the objective function continuous, and a new algorithm involving the Newton-Armijo algorithm proposed to solve the SESVR converge globally to the solution. Theoretically, we give a brief convergence proof to our algorithm. The advantages of our presented algorithm are that we only need to solve a system of linear equations iteratively instead of solving a convex quadratic program, as is the case with a conventional SVR, and lessen the influence of the penalty parameter C in our model. In order to show the efficiency of our algorithm, we employ it to forecast an actual electricity power short-term load. The experimental results show that the presented algorithm, SESVR, plays better precisely and effectively than SVM$^{\text{light}}$ and LIBSVR in stochastic time series forecasting.

Keywords: Support Vector Machine; SSVR; Entropy Function; Asymptotic Solution; Forecasting

1. Introduction

Support vector machines (SVMs), based on statistical learning theory, are powerful tools for pattern classifications and regression problems [1,2] and have been employed in engineering practices [3,4]. The basic SVM model, maximal margin classifier, needs to solve a constrained optimization mathematical programming, *i.e.*, seek for the hyperplane (w,b) to realizes the maximal margin hyperplane with geometric margin [2]:

$$\min_{(w,b)\in R^{n+1}} \frac{1}{2}\langle w\cdot w\rangle \tag{1}$$

subject to $y_i\left(\langle w,x_i\rangle+b\right)\geq 1$, $i=1,2,\cdots,l$.

For a given linearly separable training sample $S=\{(x_1,y_1),(x_2,y_2),\cdots(x_l,y_l)\}$. Generally, a simple and direct method to solve the above SVM model is to transform this optimization problem into its corresponding dual model with some constraint relations as a Lagrangian problem. Traditionally, the researchers usually transfer a constraint optimization problem into unconstrained problems to deal with SVM problems. However, after transformation, the corresponding unconstrained problem with an important plus function x_+ is not differentiable, so we can not use the traditional fast Newton method to directly solve. Fortunately, smoothing methods have been extensively used for solving important mathematical programming problems [5-8]. So, it is natural that we used the smoothing method to deal with SVM problems. The nature of the smoothing method is to construct a continue polynomial to substitute the plus function x_+ [9-15].

Chen and Mangasarian applied the penalty function to solve the SVM [16]. It is well known that the exact penalty function is better than inexact penalty function for the constrained optimization problems [17]. Meanwhile, Smooth support vector regression (SSVR) is seldom researched except Lee and Wang [18]. So, in this paper, we propose a new smooth entropy support vector regression (SESVR) model using an exact penalty function which is different from the approximating function in [16], and study its asymptotic solution approaching the solution of primal problem.

The proposed SESVR model has strong mathematical properties, such as strong convexity and infinitely often differentiability. To demonstrate the proposed SESVR's capability in solving regression problems, we employ SESVR to forecast on power short-term load forecasting from the actual electric network. We also compared our SESVR model with SVM$^{\text{light}}$ [19] and LIBSVM [20] in the aspect of forecast accuracies.

This paper is organized as follows. Standard SVM and SSVM are briefly reviewed in Section 2. In Section 3, a novel SESVR model is proposed. In Section 4, we analyze the asymptotic solution of SESVR. We use the synthetic data and actual electric power load to test the proposed SESVR and give a brief analysis in Section 5. Finally, some conclusions are drawn in Section 6.

2. A Brief Review of SSVM

2.1. Standard SVM

Given a training sample set $\{(x_i, y_i), i = 1, 2, \cdots, l\}$ with size of *l*, x_i is a column vector, $y_i = \pm 1$. The learning objective is to construct a hyperplane to correctly classify the test samples. $wx + b = \pm 1$ represents the classification hyperplane with two different data classes, the sign of data is determined by the following equations:

$$wx_i + b \ge +1,\ y_i = 1;\ wx_i + b \le -1,\ y_i = -1$$

SVM is mainly used to construct a classification hyperplane to separate two different kind samples and maximize the separation margin. For the nonlinear problem, it is necessary to introduce penalty parameter *C* and nonnegative slack variable ξ, The larger *C* is, the more severe penalty is. Therefore, the quadratic optimization problem can be obtained as following:

$$\min_{(w,b)\in R^{n+1}} \frac{1}{2}\langle w, w\rangle + C\sum_{i=1}^{l}\xi_i \qquad (2)$$

subject to

$$y_i\left(\langle w\cdot x_i\rangle + b\right) \ge 1 - \xi_i,\ i = 1, 2, \cdots, l.$$
$$\xi_i \ge 0,\ i = 1, 2, \cdots, l.$$

2.2. Smooth Support Vector Machine (SSVM)

Given a database consisting of *m* points in the n-dimensional real space R^n, which are represented by a $m \times n$ matrix, where the i^{th} row of the matrix *A* corresponds to the i^{th} data point. Two class data A_+ and A_- belong to positive (+1) and negative (−1), respectively. A $m \times n$ diagonal matrix *D* with ones or negative ones along its diagonal can be used to specify the membership of each point. In other words, $D_{ii} = \pm 1$ depending on whether the label of i^{th} data point is +1 or −1.

Combining the two constraint conditions of problem (2), we obtain

$$\begin{aligned}\xi_i &= \max\left(0, 1 - y_i\left(\langle w\cdot x_i\rangle + b\right)\right) \\ &= \left(1 - y_i\left(\langle w\cdot x_i\rangle + b\right)\right)_+\end{aligned} \qquad (3)$$

where $(\cdot)_+$ is a plus function defined as follows

$$x_+ = \begin{cases} x, x \ge 0 \\ 0, x < 0\end{cases}.$$

We substitute Equation (3) into Equation (2), and convert Equation (2) into an equivalent SVM (4) which is an unconstrained optimization problem.

$$\min_{(w,b)\in R^{n+1}} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{l}\left(\left(1 - y_i\left(\langle w\cdot x_i\rangle + b\right)\right)_+\right) \qquad (4)$$

The above function (4) has a unique solution. Function $(\cdot)_+$ is not differentiable and unsmooth, therefore, it cannot be solved using conventional Newton optimization method, because it always requires the objective function's gradient and Hessian matrix exist. Lee and Mangasarian modified the second part of function (4) and made it smooth to build a smooth unconstrained problem similar to unsmooth unconstrained problem. To do that, Lee introduced an approximation of an unsmooth function $(\cdot)_+$, which is the integral function of Sigmoid function[9,16]:

$$p(x,\alpha) = x + \frac{1}{\alpha}\log\left(1 + e^{-\alpha x}\right), \alpha > 0 \qquad (5)$$

Obviously, $p(x,\alpha)$ approach the plus function x_+ as α tends to infinity, therefore, the unconstrained optimization problem (4) is equivalent to the following smooth support vector machine (SSVM) optimization problem (6):

$$\min_{(w,b)\in R^{n+1}} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{l} p\left(\left(1 - y_i\left(\langle w\cdot x_i\rangle + b\right)\right), \alpha\right) \qquad (6)$$

The simple lemma that bounds the square difference between the plus function x_+ and its smooth approximation $p(x,\alpha)$.

Lemma 2.1 [9] For $x \in R$ and $|x| < \rho$,

$p(x,\alpha)^2 - x_+^2 < \left(\frac{\log 2}{\alpha}\right)^2 + \frac{2\rho}{\alpha}\ln 2$, where $p(x,\alpha)$ is *p* function of (5) with smoothing parameter $\alpha > 0$.

As the smoothing parameter *α* approaches infinity, the unique solution of our smooth problem (6) using Newton-Armijo Algorithm approaches the unique solution of the equivalent SVM problem (2).

Theorem 2.2: [9] Let $\{(w^i, b^i)\}$ be a sequence generated by Newton-Armijo Algorithm and $(\overline{w}, \overline{b})$ be the unique solution of problem (6).

1) The sequence (w^i, b^i) converges to the unique solution ($\overline{w}, \overline{b}$) from any initial point (w^0, b^0) in R^{n+1}.

2) For any initial point (w^0, b^0), there exists an integer $\overline{i}$ such that the stepsize λ_i of Newton-Armijo Algorithm equals 1 for $i \ge \overline{i}$ and the sequence $\{(w^i, b^i)\}$ converges to $(\overline{w}, \overline{b})$.

Recently, some smooth functions are constructed to replace the plus function x_+, for example, tangent circular arc smooth piecewise function in [11], cubic spline interpolation function and Hermite interpolation polynomial in [12], two piecewise smooth functions (1PSSVM, 2PSSVM) in [13] and so on.

3. The Proposed SESVR Model

Given a data set *S* which consists of *l* points in *n*-dimensional real space R^n and l observations of real value associated with each point, that is,

$$S=\{(x_i,y_i),x_i\in R^n,y_i\in R,i=1,2,\cdots,l\}$$

We would like to find a linear or nonlinear regression function, $f(x)$, tolerating a small error in fitting this given data set. This can be achieved by utilizing the ε-insensitive loss function that sets an ε-insensitive "tube" around the data, within which errors are discarded. Also, applying the idea of support vector machines (SVMs) [2]. The function $f(x)$ is made as flat as possible in fitting the training data set. The idea of representing the solution by means of a small subset of training points has enormous computational advantages. The ε-insensitive loss function maintains that advantages, while still ensuring the existence of a global minimum and the optimization of a reliable generalization bound.

This linear regression problem can be formulated as an unconstrained minimization problem given as follows:

$$\min_{(w,b)\in R^{n+1}}\frac{1}{2}\|w\|_2^2+C\cdot 1^T|\xi|_\varepsilon \tag{7}$$

where, $(|\xi|_\varepsilon)_i=\max(0,|w\cdot x_i+b-y_i|-\varepsilon)$, $|\xi|_\varepsilon\in R^l$.

That represents the fitting errors and the positive control parameter C here weights the tradeoff between the fitting errors and the flatness of the linear regression function $f(x)$. To deal with the ε-insensitive loss function in the objective function of the above minimization problem, conventionally, it is reformulated as a constrained minimization problem defined as follows:

$$\min_{(w,b,\xi,\xi^*)\in R^{n+1+2l}}\frac{1}{2}\langle w,w\rangle+C\sum_{i=1}^{l}(\xi_i+\hat{\xi}_i)$$

Subject to

$$\begin{aligned}&(\langle w\cdot x_i\rangle+b)-y_i\le\varepsilon+\xi_i\\&y_i-(\langle w\cdot x_i\rangle+b)\le\varepsilon+\hat{\xi}_i\\&\xi_i,\hat{\xi}_i\ge 0,\ i=1,2,\cdots,l\end{aligned}\tag{8}$$

This formulation (8), which is equivalent to the formulation (7), is a convex quadratic minimization problem with $n+1$ free variables, 2 l nonnegative variables, and 2 l inequality constraints. However, more variables and constraints in the formulation enlarges the problem size and could increase computational complexity for solving the regression problem.

We denote

$$\begin{aligned}&\xi_i^+(w,b)\\&=\max(0,(\langle w\cdot x_i\rangle+b)-y_i-\varepsilon,y_i-(\langle w\cdot x_i\rangle+b)-\varepsilon)\end{aligned}$$

$$\xi^+(w,b)=(\xi_1^+(w,b),\xi_2^+(w,b),\cdots,\xi_{2l}^+(w,b))$$

and $i=1,2,\cdots,2l$. Adopting Fletcher's nonlinearly undifferentiable precise penalty method [21], we obtain the equivalent unconstraint optimization problem:

$$\min_{(w,b)\in R^{n+1}}\phi(w,b)=\frac{1}{2}\langle w,w\rangle+C\|\xi^+(w,b)\|_1 \tag{9}$$

Lemma 3.1 [22] (w^*,b^*) is an arbitrary point in R^{n+1}, λ^* is a Lagrange multiplier vector. If parameter C satisfies $C\ge\frac{1}{2}\|\lambda^*\|_\infty$ and $\xi^+(w^*,b^*)=0$. Then the 2nd order optimal sufficient condition is equivalent between the optimization problem (9) and problem (8) in the point (w^*,b^*). where $\|\cdot\|_\infty$ is a dual norm of $\|\cdot\|_1$.

In term of Lemma 3.1, inter-dual norm $\|\cdot\|_\infty$ and $\|\cdot\|_1$, when the trade-off penalty factor C satisfying $C>\|\lambda^*\|_1$, we can solve the following optimal problem:

$$\begin{aligned}&\min_{(w,b)\in R^{n+1}}\phi(w,b):=\frac{1}{2}\langle w,w\rangle\\&+C\cdot\max\{0,\xi_1(w,b),\cdots\xi_{2l}(w,b)\}\end{aligned}\tag{10}$$

and obtain a feasible solution of the optimization problem (9). However, maximum function $\max\{0,\xi_1(w,b),\cdots\xi_{2l}(w,b)\}$ is not smooth and undifferentiable, so that, we can not directly utilize Newton gradient descent method. We employ the following smooth entropy function $p_r(\xi_1,\cdots\xi_{2l})=r\ln\left\{1+\sum_{i=1}^{2l}\exp(\xi_i(w,b)/r)\right\}$ and substitute it for approximating maximum function $\max\{0,\xi_1(w,b),\cdots\xi_{2l}(w,b)\}$.

Lemma 3.2: $\lim_{r\to 0}p_r(\xi_1,\xi_2,\cdots\xi_{2l})=\max\{0,\xi_1,\xi_2,\cdots\xi_{2l}\}$ for arbitrary $\xi_1,\xi_2,\cdots\xi_{2l}$.

Then, we obtain the smooth and differentiable optimization model (11) which is equivalent to problem (9):

$$\begin{aligned}&\min_{(w,b)\in R^{n+2}}\phi_r(w,b):=\frac{1}{2}\|w\|_2^2\\&+C\cdot r\ln\left\{1+\sum_{i=1}^{2l}\exp(\xi_i(w,b)/r)\right\}\end{aligned}\tag{11}$$

We call the above problem (11) a smooth entropy support vector regression (SESVR).

4. The Asymptotic Analysis of SESVR

The proposed SESVR problem (11) is an infinitely differentiable optimization issue, and can be solved using Newton iterating algorithm. We hope solve the problem (11) instead of the problem (9), however, so far, we don't know the solution relations between the problem (11) and the primal SVR problem (8). In fact, when $C>\|\lambda^*\|_1$, although the solution of the primal problem (8) is not the optimal solution of the problem (11), we can prove that the convergence bound between the solution of the problem (11) and the solution of the problem (9) is controlled by smoothing parameter *r i.e*:

Theorem 4.1: Suppose that $\left(w^*,b^*\right)$ is a optimal solution of problem (9), and λ^* is Lagrange multiplier vector, if $C>\left\|\lambda^*\right\|_1$, then for an arbitrary point $\left(w,b\right)\in R^{n+1}$ and a given small $r>0$, we have

$$\phi_r\left(w^*,b^*\right)\le\phi_r\left(w,b\right)+C\cdot r\ln\left(2l+1\right) \qquad (12)$$

Proof: Because point $\left(w^*,b^*\right)$ is the optimal solution of problem (9), and λ^* is Lagrange multiplier vector, we obtain the Karush-Kuhn-Tucher complementarity conditions:

$$\begin{aligned}&\lambda_i^*\xi_i\left(w^*,b^*\right)=0,\lambda_i^*\ge 0,\xi_i\left(w^*,b^*\right)\ge 0\\&i=1,2,\cdots,m.\end{aligned} \qquad (13)$$

From Lemma 3.1, we know $\left(w^*,b^*\right)$ is the optimal solution of problem (10), so, we obtain inequality (14)

$$\begin{aligned}&\phi\left(w,b\right)\ge\phi\left(w^*,b^*\right)\\&+C\cdot\max\left\{0,\xi_1\left(w^*,b^*\right),\cdots,\xi_{2l}\left(w^*,b^*\right)\right\}+\sum_{i=1}^{2l}\lambda_i^*\xi_i\left(w^*,b^*\right)\\&\ge\phi\left(w^*,b^*\right)\\&+\sum_{i=1}^{2l}\lambda_i^*\left(\max\left\{0,\xi_1\left(w^*,b^*\right),\cdots,\xi_{2l}\left(w^*,b^*\right)\right\}-\xi_i\left(w^*,b^*\right)\right)\\&\ge\phi\left(\left(w^*,b^*\right)\right)\end{aligned} \qquad (14)$$

On the other hand, for any given point $\left(w,b\right)\in R^{n+1}$, penalty parameter $C>0$, and arbitrary small constant $r>0$, we have

$$\phi\left(w,b\right)\le\phi_r\left(w,b\right)\le\phi\left(w,b\right)+C\cdot r\ln\left(2l+1\right) \qquad (15)$$

Considering $\left(w^*,b^*\right)$ being a feasible solution, then we get

$$\phi_r\left(w^*,b^*\right)\le\phi\left(w^*,b^*\right)+C\cdot r\ln\left(2l+1\right) \qquad (16)$$

From Equations (14), (15), and (16), we know

$$\begin{aligned}\phi_r\left(w^*,b^*\right)&\le\phi\left(w^*,b^*\right)+C\cdot r\ln\left(2l+1\right)\\&\le\phi\left(w,b\right)+C\cdot r\ln\left(2l+1\right)\\&\le\phi_r\left(w,b\right)+C\cdot r\ln\left(2l+1\right)\end{aligned}$$

Then, the theorem 4.1 is correct.

Theorem 4.2: Suppose that $\left(w^*,b^*\right)$ is an optimal solution of problem (9), λ^* is Lagrange multiplier vector, and $\left(\overline{w},\overline{b}\right)$ is the optimal solution of problem (11), if $C>\left\|\lambda^*\right\|_1$, then

$$\begin{aligned}-C\cdot r\ln\left(2l+1\right)&\le\phi\left(w^*,b^*\right)-\phi\left(\overline{w},\overline{b}\right)\\&\le 3C\cdot r\ln\left(2l+1\right)\end{aligned} \qquad (17)$$

The proof is similar to Theorem 3.4 in [22], so we abandon the redundant proof.

Theorem 4.2 shows that the solution of SESVR (11) approaches the solution of primal problem (8) as the smoothing parameter $r\to 0$. By making use of this results and taking advantage of the twice differentiability of the objective function, we prescribe a globally convergent algorithm 4.3 based on Newton-Armijo algorithm for solving (11) as follows.

Algorithm 4.3: Start with any choice of initial point $\left(w_0,b_0\right)\in R^{n+1}$, and stop if $\left(w_i,b_i\right)$ satisfy $\left\|w_{i+1}-w_i\right\|\le\delta$ and $\left\|b_{i+1}-b_i\right\|\le\delta$ for a given sufficiently small constant δ.

Step 1: Initialize $C_0=1$, $r_0=1$.

Step 2: For $k=1,2,\cdots$, using Newton-Armijo algorithm [9,16], we solve the unconstraint optimization problem:

$$\left(w_k,b_k\right)\in\arg\min_{(w,b)\in R^{n+1}}\phi_{r_k}\left(w,b\right)$$

Step 3: Let $r_{k+1}:=r_k/2$. If $\left(w_k,b_k\right)$ is a feasible solution of (11), then $C_{k+1}:=C_k$, otherwise, $C_{k+1}:=2C_k$.

From Algorithm 4.3, we can easily validate the following facts.

1) $r_k\to 0$ and

$p_{r_k}\left(\xi_1,\cdots,\xi_{2l}\right)\to\max\left\{0,\xi_1\left(w,b\right),\cdots\xi_{2l}\left(w,b\right)\right\}$ as $k\to\infty$.

2) $C_k\cdot r_k<1$ and for an arbitrary $\left(\tilde{w},\tilde{b}\right)$,

$$\begin{aligned}&C_kp_{r_k}\left(\left(\xi_1,\cdots,\xi_{2l}\right)\right)\\&=C_kr_k\ln\left\{1+\sum_{i=1}^{2l}\exp\left(\xi_i\left(\tilde{w},\tilde{b}\right)/r_k\right)\right\}\to 0.\end{aligned}$$

3) If the sequences $\left\{\left(w_k,b_k\right)\right\}$ contain a non-feasible infinite sub-sequence, then $r_k\to\infty$.

With finitely iterating, the sequence $\left\{\left(w_k,b_k\right)\right\}$ of Algorithm 4.3 globally converge to the unique solution based on the following Lemma 4.4 - 4.5 and Theorem 4.6.

Lemma 4.4: The sequence $\left\{\left(w_k,b_k\right)\right\}$ is boundary and all the corresponding cluster points are feasible points of problem (9).

Lemma 4.5: Any cluster point of sequence $\left\{\left(w_k,b_k\right)\right\}$ is the optimal solution of problem (9).

Theorem 4.6: Let {(w_k, b_k)} be a sequence generated by Algorithm 4.3 and ($\overline{w},\overline{b}$) be the unique solution of problem (9).

1) The sequences $\left\{\left(w_k,b_k\right)\right\}$ converge to unique solution $\left(\overline{w},\overline{b}\right)$ from any initial point $\left(w_0,b_0\right)\in R^{n+1}$.

2) For any initial point $\left(w_0,b_0\right)$, there exists an integer $\overline{k}$ such that the stepsize λ_i of Newton-Armijo Algorithm equals 1 for $k\ge\overline{k}$ and the sequence {(w_k, b_k)} converges to $\left(\overline{w},\overline{b}\right)$.

Lemma 4.5, Lemma 4.6 and Theorem 4.6 can be inferred from [9,18]. In the above discussion, we construct the smooth entropy support vector regression model for a linear regression function in fitting the given training data points under the criterion that minimizes the squares of the ε-insensitive loss function. That is approximating $y \in R^{2l}$ by a linear function of the form:

$$y = w^T \cdot x + 1^T b \tag{18}$$

where $w \in R^n$ and $b \in R$ are parameters to be determined by minimizing the objective function in (11). Applying the duality theorem in convex minimization problem [23], w can be represented by $A^T u$ for some $u \in R^{2l}$. Hence, we have

$$y = AA^T u + 1^T b \tag{19}$$

This motivated the nonlinear support vector regression model. In order to generalize our results from the linear case to nonlinear case, we employ the kernel technique that has been used extensively in kernel-based learning algorithms [1,2].

We simply replace the AA^T in (19) by a nonlinear kernel matrix $K(A, A^T)$, where $K\left(A, A^T\right)_{i,j} = K\left(A_i, A_j^T\right)$ and $K\left(x^T, z\right)$ is a nonlinear kernel function. Using the same loss criterion with the linear case, this will give us the nonlinear support vector regression formulation as follows:

$$\min_{(w,b)\in R^{n+1}} \frac{1}{2}u^T AA^T u + \frac{C}{2}\sum_{i=1}^{2l} \xi_i\left(u,b\right) + \hat{\xi}_i\left(u,b\right) \tag{20}$$

where

$$\xi_i\left(u,b\right) = \max\left(0, K\left(A_i, A^T\right)u + b - y_i - \varepsilon\right),$$

$$\hat{\xi}_i\left(u,b\right) = \max\left(0, y_i - K\left(A_i, A^T\right)u - b + \varepsilon\right),$$

$$i = 1,2,\cdots,l.$$

$K\left(A_i, A^T\right)$ is a kernel map from $R^{1\times n} \times R^{n\times 2l}$ to $R^{1\times 2l}$. We solve the optimal solution (20) using Algorithm 4.3, and obtain parameters u and b. Then the regression function is

$$\begin{aligned} y &= K\left(x^T, A^T\right)u + b = u^T K\left(A, x\right) + b \\ &= \sum_{i=1}^{2l} u_i K\left(A_i, x\right) + b \end{aligned} \tag{21}$$

5. Experimental Results and Analysis

In order to test the efficiency of our proposed smooth entropy support vector regressions, we utilize SESVR to forecast electricity power short term load and compared the results with the conventional SVM$^{\text{light}}$ [19] and LIBSVM [20].

All experiments were run on a personal computer, which consisted of a 1.9 GHz AMD dual core processor and 960 megabytes of memory. Based on the first order optimality conditions of unconstrained convex minimization problem, our stop criterion for the proposed model was satisfied when the gradient of the objective function is less than 10^{-5} and select $\delta = 10^{-4}$. We implemented SESVR in C++ programming. In the experiments, 2-norm relative error was chosen to evaluate the tolerance between the predicted values and the actual values. For an actual value y and the predicted vector $\hat{y}$, the 1-norm relative error of two vectors y and $\hat{y}$ was defined as follows:

$$E = \frac{1}{2l}\sum_{i=1}^{2l}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{22}$$

Aim to evaluate how well each method generalized to unseen data, we split the entire data set into two parts, the training set and testing set. The training data was used to generate the regression function that is learning from training data; the testing set, which is not involved in the training procedure, was used to evaluate the prediction ability of the resulting regression function. A smaller testing error indicates a better prediction ability. We performed tenfold cross-validation on each data set and reported the average testing error in our numerical results.

Our actual power load data and its corresponding weather data come from HuaiNan electric network from April 2004 to August 2004, sampling frequency is one times each hour, sampling days is 120, so we gain 2880 (120 × 24) training data. We use trained SESVR model to forecast the load of August 10 2005. It is know that summer season influence power load more drastically than other three seasons, it is why we sample load from summer season, meanwhile, we can study whether or not weather influence and how to influence electricity load. The numerical results of short-term load forecasting were also included in following **Table 1**.

From **Table 1** and **Figure 1**, we can find out our proposed SESVR model is a feasible forecasting method for electric power short-term load, If we generated the training samples for SESVR including the weather temperatures, the relative error is 1.28%, otherwise the error is 3.06%. So, the weather temperatures can upgrade forecasting accuracies, moreover, we can see that the relative errors of night time is bigger than that of daytime. On the other hand, we find out the penalty parameter *C* in Algorithm 4.3 increases monotonously, so Algorithm 4.3 is stable, however, SVM$^{\text{light}}$ and LIBSVM are influenced by penalty parameter *C*. In order to gain the optimal solutions of SVM$^{\text{light}}$ and LIBSVM, we must tune *C* carefully. This increases the computational complexity. With regard to this point, Algorithm 4.3 is superior to SVM$^{\text{light}}$ and LIBSVM.

Figure 2 illustrates the tenfold numerical results and comparisons of our proposed SESVR, SVM, LIBSVM. The experimental results demonstrated that our proposed SESVR model is a powerful tool for forecast electric power short-term load, and better precise and effective

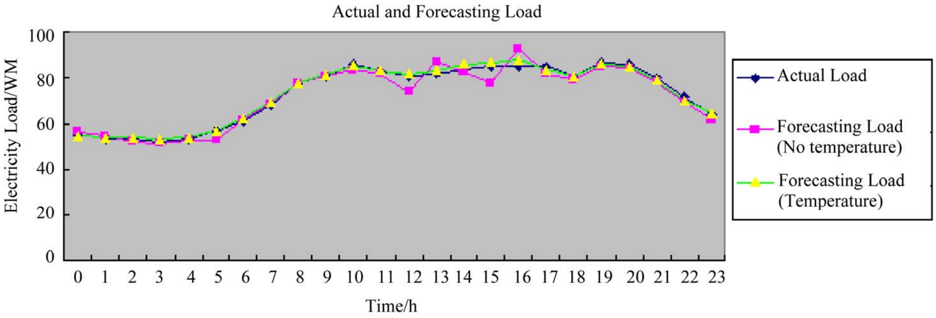

Figure 1. The electricity load forecasting results using our proposed SESVR model. Blue line, red line, and green line show the actual load, forecasting load with no temperature, and forecasting load with temperature, respectively.

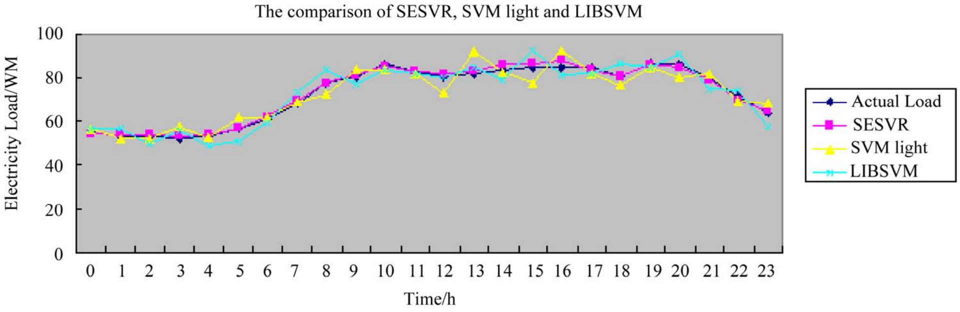

Figure 2. The electricity load forecasting comparison of SESVR, SVM^{light}, LIBSVM. Blue line shows the actual load, read line represents the forecasting load using our proposed regression model SESVR, yellow line and green line illustrate the forecasting loads using SVM^{light} and LIBSVM model, respectively.

Table 1. The load forecasting result on 10 August 2005.

Time/h	Actual load /MW	No temp.	Temp.
		Forecasting /MW	Forecasting /MW
0	55	56.20	54.79
1	53	54.30	53.80
2	53	52.06	53.67
3	52	51.75	53.12
4	53	52.56	53.66
5	57	52.40	57.12
6	61	61.59	62.08
7	68	68.93	69.33
8	78	77.98	77.79
9	81	80.79	81.31
10	86	84.13	85.42
11	83	81.96	83.05
12	81	74.30	81.78
13	82	87.20	83.19
14	84	82.87	86.09
15	85	77.80	86.71
16	85	92.70	87.89
17	85	81.85	84.01
18	81	79.51	80.98
19	87	84.93	86.36
20	86	84.62	84.83
21	80	78.38	78.95
22	72	69.71	69.83
23	64	61.86	64.59
Relative error %		3.06%	1.28%

than LIBSVM and SVM^{light} model.

6. Conclusions

We have proposed a novel formulation, SESVR, which is a smooth unconstrained optimization reformulation of the traditional linear programming associated with an ε-I nsensitive support vector regression. We have employed an entropy penalty function to substitute it for the plus function to avoid objective discontinuous. We have proposed a new algorithm involving a very fast Newton-Armijo algorithm to solve the SESVR that has been shown convergent globally to the solution. In our algorithm, the penalty parameter *C* is creasing monotonously, the influence to SESVR performance is smaller than foregoing SVM^{light} and LIBSVM. Theoretically, we have given a brief convergence proof to our algorithm.

In order to show the efficiency of our algorithm, we employ it to forecast an actual electricity power short-term load. The experimental results show that the proposed SESVR is effective and precise, and plays better performances than SVM^{light} and LIBSVR in stochastic time series forecasting. Moreover, an advantage of our proposed SESVR algorithm is that we only need to solve a system of linear equations iteratively instead of solving a convex quadratic program, as is the case with a conventional SVR.

REFERENCES

[1] C. V. Gustavo, G. Juan and G. P. Gabriel, “On the Suitable Domain for SVM Training in Image Coding,” *Journal of Machine Learning Research*, Vol. 9, No. 1, 2008, pp. 49-66.

[2] F. Chang, C. Y. Guo and X. R. Lin, “Tree Decomposition for Large-Scale SVM Problems,” *Journal of Machine Learning Research*, Vol. 11, No. 10, 2010, pp. 2935-2972.

[3] Y. H. Kong, W. C. Wei and W. He, “Power Quality Disturbance Signal Classification Using Support Vector Machine Based on Feature Combination,” *Journal of North China Electric Power University*, Vol. 37, No. 4, 2010, pp. 72-77.

[4] J. Zhe, “Research on Power Load Forecasting Base on Support Vector Machines,” *Computer Simulation*, No. 8, 2010, pp. 282-285.

[5] B. Chen and P. T. Harker, “Smooth Approximations to Nonlinear Complementarity Problems,” *SIAM Journal of Optimization*, Vol. 7, No. 2, 1997, pp. 403-420. doi:10.1137/S1052623495280615

[6] C. H. Chen and O. L. Mangasarian, “Smoothing Methods for Convex Inequalities and Linear Complementarity Problems,” *Mathematical Programming*, Vol. 71, No. 1, 1995, pp. 51-69. doi:10.1007/BF01592244

[7] C. H. Chen and O. L. Mangasarian, “A Class of Smoothing Functions for Nonlinear and Mixed Complementarity Problems,” *Computational Optimization and Applications*, Vol. 5, No. 2, 1996, pp. 97-138. doi:10.1007/BF00249052

[8] X. Chen, L. Qi and D. Sun, “Global and Superlinear Convergence of the Smoothing Newton Method and Its Application to General Box Constrained Variational Inequalities,” *Mathematics of Computation*, Vol. 67, No. 222, 1998, pp. 519-540. doi:10.1090/S0025-5718-98-00932-6

[9] Y. J. Lee and O. L. Mangasarin, “SSVM: A Smooth Support Vector Machine for Classification,” *Computational Optimization and Applications*, Vol. 20, No. 1, 2010, pp. 5-22. doi:10.1023/A:1011215321374

[10] Z. Q. Meng, G. G. Zhou and Y. H. Zhu, “A Smoothing Support Vector Machine Based on Exact Penalty Function,” *Lecture Notes in Artificial Intelligence*, Vol. 3801, 2005, pp. 568-573.

[11] Y. F. Fan, D. X. Zhang and H. C. He, “Tangent Circular Arc Smooth SVM（TCA-SSVM） Research,” 2008 *Congress on Image and Signal Processing*, Sanya, 27-30 May 2008, pp. 646-648. doi:10.1109/CISP.2008.112

[12] Y. F. Fan, D. X. Zhang and H. C. He, “Smooth SVM Research: A Polynomial-Based Approach,” *The 9th International Conference on Information and Communications Security*, Singapore, 10-13 December 2007, pp. 983-988.

[13] L. K. Lao, C. D. Lin, H. Peng, *et al.*, “A Study on Piecewise Polynomial Smooth Approximation to the Plus Function,” *The 9th International Conference on Control*, *Automation*, *Robotics and Vision*, Singapore, 5-8 December 2006, pp. 342-347.

[14] J. Z. Xiong, T. M. Hu and G. G. Li, “A Comparative Study of Three Smooth SVM Classifiers,” *Proceedings of the 6th World Congress on Intelligent Control and Automation*, Dalian, 21-23 June 2006, pp. 5962-5966.

[15] P. A. Forero, A. C. Georgios and B. Giannakis, “Consensus-Based Distributed Support Vector Machines,” *Journal of Machine Learning Research*, Vol. 11, No. 5, 2010, pp. 1663-1707.

[16] C. H. Chen and O. L. Mangasarian, “Smoothing Methods for Convex Inequalities and Linear Complementarity Problems,” *Mathematical Programming*, Vol. 71, No. 1, 1995, pp. 51-69. doi:10.1007/BF01592244

[17] S. H. Peng and X. S. Li, “The Asymptotic Analysis of Quasi-Exact Penalty Function Method,” *Journal of Computational Mathematics*, Vol. 29, No. 1, 2007, pp. 47-56.

[18] Y. J. Lee, W. F. Hsieh and C. M. Huang, “ε-SSVR: A Smooth Support Vector Machine for ε-Insensitive Regression,” *Knowledge and Data Engineering*, Vol. 17, No. 5, 2005, pp. 678-695. doi:10.1109/TKDE.2005.77

[19] T. Joachims, “SVM$^{\text{light}}$,” 2010.
http://svmlight. joachims .org

[20] C.-C. Chang and C.-J. Lin, “LIBSVM: A Library for Support Vector Machines,” 2010.
http://www.csie.ntu. edu.tw/~cjlin/libsvm

[21] R. Fletcher, “Practical Methods of Optimization,” John Wiley & Sons, New York, 1981.

[22] O. L. Mangasarian and D. R. Musicant, “Successive Overrelaxation for Support Vector Machines,” *IEEE Transactions on Neural Networks*, Vol. 10, No. 5, 2010, pp. 1032-1037. doi:10.1109/72.788643
ftp://ftp.cs.wisc.edu/math-prog/tech-reports/98-18.ps

[23] Y. W. Chang, C. J. Hsieh and K. W. Chang, “Training and Testing Low-degree Polynomial Data Mappings via Linear SVM,” *Journal of Machine Learning Research*, Vol. 11, No. 4, 2010, pp. 1471-1490.