Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 243921, 8 pages
http://dx.doi.org/10.1155/2014/243921

Research Article

Analyzing Big Data with the Hybrid Interval Regression Methods

Chia-Hui Huang,[1] Keng-Chieh Yang,[2] and Han-Ying Kao[3]

[1] Department of Business Administration, National Taipei University of Business, No. 321, Section 1, Jinan Road, Zhongzheng District, Taipei City 100, Taiwan
[2] Department of Information Management, Hwa Hsia Institute of Technology, No. 111, Gongzhuan Road, Zhonghe District, New Taipei City 235, Taiwan
[3] Department of Computer Science and Information Engineering, National Dong Hwa University, No. 123, Hua-Shi Road, Hualien 97063, Taiwan

Correspondence should be addressed to Chia-Hui Huang; leohkkimo@gmail.com

Received 19 May 2014; Accepted 7 July 2014; Published 20 July 2014

Academic Editor: Jung-Fa Tsai

Big data is a new trend at present, forcing the significant impacts on information technologies. In big data applications, one of the most concerned issues is dealing with large-scale data sets that often require computation resources provided by public cloud services. How to analyze big data efficiently becomes a big challenge. In this paper, we collaborate interval regression with the smooth support vector machine (SSVM) to analyze big data. Recently, the smooth support vector machine (SSVM) was proposed as an alternative of the standard SVM that has been proved more efficient than the traditional SVM in processing large-scale data. In addition the soft margin method is proposed to modify the excursion of separation margin and to be effective in the gray zone that the distribution of data becomes hard to be described and the separation margin between classes.

1. Introduction

Big data has become one of new research frontiers. Generally speaking, big data is a collection of large-scale and complex data sets that it becomes more difficult to process using current database management systems and traditional data processing applications. In 2012, Gartner Inc. gave a definition of big data as "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization" [1]. The trend of big data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data.

One of the major applications of the future parallel, distributed, and cloud systems is in big data analytic [2–5]. Most concerned issues are dealing with large-scale sets which often require computation resources provided by public cloud services. How to analyze big data efficiently becomes a big challenge.

The support vector machine (SVM) has shown to be an efficient approach for a variety of data mining, classification, analysis, pattern recognition, and distribution estimation [6–14]. Recently, using SVM to solve the interval regression model [15] has become an alternative approach. Hong and Hwang [16] evaluated interval regression models with quadratic loss SVM. Bisserier et al. [17] proposed a revisited fuzzy regression method where a linear model is identified from Crisp-Inputs Fuzzy-Outputs (CISO) data. D'Urso et al. [18] presented fuzzy clusterwise regression analysis with LR fuzzy response variable and numeric explanatory variables. The suggested model is to allow for linear and nonlinear relationship between the output and input variables. Jeng et al. [19] developed a support vector interval regression networks (SVIRNs) based on both SVM and neural networks. Huang and Kao [20] proposed a soft-margin SVM for interval regression analysis. Huang [21] solved interval regression model with reduced support vector machine.

However, there are several main problems while using SVM model.

(1) Big data: when dealing with big data sets, the solution by using SVM with a nonlinear kernel may be difficult to be found.

(2) Noises and interaction: the distribution of data becomes hard to be described and the separation margin between classes becomes a "gray" zone.

(3) Unbalance: the number of samples from one class is much larger than the number of samples from other classes. It causes the excursion of separation margin.

Under this circumstance, developing an efficient method to analyze big data becomes important. The smooth support vector machine (SSVM) has been proved more efficient than the traditional SVM in processing large-scale data [22–24]. The main idea of SSVM is solved by a fast Newton-Armijo algorithm [25] and has been extended to nonlinear separation surfaces by using a nonlinear kernel technology [24].

In this study, we collaborate interval regression [15] with SSVM to analyze big data. The main idea of SSVM is solved by a fast Newton-Armijo algorithm and has been extended to nonlinear separation surfaces by using a nonlinear kernel technology. Additionally, to modify the excursion of separation margin and to be effective in the gray zone, the soft margin method is proposed. The experiment results show that the proposed methods are more efficient than existing methods.

This study is organized as follows. Section 2 reviews the current methods for interval regression analysis. Section 3 proposes the soft margin method and the formulation of interval regression with SSVM to analyze big data. Section 4 gives a numerical example by the proposed methods dealing with big data which is extracted from Taiwan Stock Exchange Capitalization Weighted Stock Index (TAIEX) [26]. Finally, Section 5 gives the concluding remarks.

2. Literature Review

Since Tanaka et al. [27] introduced the fuzzy regression model with symmetric fuzzy parameters, the properties of fuzzy regression have been studied extensively by many researchers. Fuzzy regression model can be simplified to interval regression analysis which is considered as the simplest version of possibilistic regression analysis with interval coefficients. An interval linear regression model is described as

$$Y\left(\mathbf{x}_j\right) = A_0 + A_1 x_{1j} + \cdots + A_n x_{nj}, \tag{1}$$

where $Y(\mathbf{x}_j)$, $j = 1, 2, \ldots, q$, is the estimated interval corresponding to the real input vector $\mathbf{x}_j = (x_{1j}, x_{2j}, \ldots, x_{nj})^t$. An interval coefficient A_i is defined as (a_i, c_i), where a_i is the center and c_i is the radius. Hence, A_i can also be represented as

$$A_i = \left[a_i - c_i, a_i + c_i\right] = \left\{a_i - c_i \le a \le a_i + c_i\right\}. \tag{2}$$

The interval linear regression model (1) can also be expressed as

$$\begin{aligned} Y\left(\mathbf{x}_j\right) &= A_0 + A_1 x_{1j} + \cdots + A_n x_{nj} \\ &= \left(a_0, c_0\right) + \left(a_1, c_1\right) x_{1j} + \cdots + \left(a_n, c_n\right) x_{nj} \\ &= \left(a_0 + \sum_{i=1}^{n} a_i x_{ij}, c_0 + \sum_{i=1}^{n} c_i \left|x_{ij}\right|\right). \end{aligned} \tag{3}$$

For a data set with crisp inputs and interval outputs, two interval regression models, the possibility and necessity models, are considered. By assumption, the center coefficients of the possibility regression model and the necessity regression model are the same [15]. For this data set, the possibility and necessity estimation models are defined as

$$\begin{aligned} Y^*\left(\mathbf{x}_j\right) &= A_0^* + A_1^* x_{1j} + \cdots + A_n^* x_{nj} \\ Y_*\left(\mathbf{x}_j\right) &= A_{0*} + A_{1*} x_{1j} + \cdots + A_{n*} x_{nj}, \end{aligned} \tag{4}$$

where the interval coefficients A_i^* and A_{i*} are defined as $A_i^* = (a_i^*, c_i^*)$ and $A_{i*} = (a_{i*}, c_{i*})$, respectively. The interval $Y^*(\mathbf{x}_j)$ estimated by the possibility model must include the observed interval Y_j and the interval $Y_*(\mathbf{x}_j)$ estimated by the necessity model must be included in the observed interval Y_j.

In this section, we review the current methods which are ordinarily used for interval regression analysis.

2.1. Tanaka and Lee's Approach. Tanaka and Lee [15] proposed an interval regression analysis with a quadratic programming (QP) approach which gives more diverse spread coefficients than a linear programming (LP) one.

The interval regression analysis by QP approach unifying the possibility and necessity models subject to the inclusion relations, $Y_*(\mathbf{x}_j) \subseteq Y_j \subseteq Y^*(\mathbf{x}_j)$, can be represented as

$$\begin{aligned} \min \quad & \sum_{j=1}^{q} \left(d_0 + \sum_{i=1}^{n} d_i \left|x_{ij}\right|\right)^2 + \varphi \sum_{i=0}^{n} \left(a_i^2 + c_i^2\right) \\ \text{s.t.} \quad & Y_*\left(\mathbf{x}_j\right) \subseteq Y_j \subseteq Y^*\left(\mathbf{x}_j\right), \quad j = 1, 2, \ldots, q \\ & c_i, d_i \ge 0, \quad i = 0, 1, \ldots, n, \end{aligned} \tag{5}$$

where φ is an extremely small positive number and makes the influence of the term $\varphi \sum_{i=0}^{n} (a_i^2 + c_i^2)$ on the objective function negligible. The constraints of the inclusion relations are equivalent to

$$
\begin{aligned}
&Y_*\left(\mathbf{x}_j\right) \subseteq Y_j \Longleftrightarrow \\
&\begin{cases}
y_j - e_j \le \left(a_0 + \sum_{i=1}^{n} a_i x_{ij}\right) - \left(c_0 + \sum_{i=1}^{n} c_i \left|x_{ij}\right|\right) \\
\left(a_0 + \sum_{i=1}^{n} a_i x_{ij}\right) + \left(c_0 + \sum_{i=1}^{n} c_i \left|x_{ij}\right|\right) \le y_j + e_j,
\end{cases}
\end{aligned}
\tag{6}
$$

$$
\begin{aligned}
&Y_j \subseteq Y^*\left(\mathbf{x}_j\right) \Longleftrightarrow \\
&\begin{cases}
\left(a_0 + \sum_{i=1}^{n} a_i x_{ij}\right) - \left(c_0 + \sum_{i=1}^{n} c_i \left|x_{ij}\right|\right) \\
\qquad - \left(d_0 + \sum_{i=1}^{n} d_i \left|x_{ij}\right|\right) \le y_j - e_j \\
y_j + e_j \le \left(a_0 + \sum_{i=1}^{n} a_i x_{ij}\right) + \left(c_0 + \sum_{i=1}^{n} c_i \left|x_{ij}\right|\right) \\
\qquad + \left(d_0 + \sum_{i=1}^{n} d_i \left|x_{ij}\right|\right),
\end{cases}
\end{aligned}
\tag{7}
$$

where $\mathbf{x}_j$ is the jth input vector and Y_j is the corresponding interval output that consists of a center y_j and a radius e_j denoted by $Y_j = (y_j, e_j)$.

2.2. Hong and Hwang's Approach. Hong and Hwang [16] evaluated interval regression model combining the possibility and necessity estimation formulation with the principle of quadratic loss support vector machine (QLSVM). This version of SVM utilizes the quadratic loss function. The QLSVM performs interval nonlinear regression analysis by constructing an interval linear regression function in high-dimensional feature space.

With the principle of QLSVM, the interval nonlinear regression model is given as follows:

$$
\begin{aligned}
\max \quad & -\frac{1}{2}\Bigg(\sum_{i,j=1}^{n}\left(\lambda_{2i}-\lambda_{2i}^*\right)\left(\lambda_{2j}-\lambda_{2j}^*\right)K\left(\mathbf{x}_i,\mathbf{x}_j\right) \\
&\quad + \sum_{i,j=1}^{n}\left(\lambda_{3i}-\lambda_{3i}^*\right)\left(\lambda_{3j}-\lambda_{3j}^*\right)K\left(\mathbf{x}_i,\mathbf{x}_j\right) \\
&\quad + \sum_{i,j=1}^{n}\left(\lambda_{4i}-\lambda_{4i}^*\right)\left(\lambda_{4j}-\lambda_{4j}^*\right)K\left(\mathbf{x}_i,\mathbf{x}_j\right) \\
&\quad + 2\sum_{i,j=1}^{n}\left(\lambda_{2i}-\lambda_{2i}^*\right)\left(\lambda_{3j}-\lambda_{3j}^*\right)K\left(\mathbf{x}_i,\mathbf{x}_j\right) \\
&\quad - 2\sum_{i,j=1}^{n}\left(\lambda_{2i}-\lambda_{2i}^*\right)\left(\lambda_{4j}-\lambda_{4j}^*\right)K\left(\mathbf{x}_i,\mathbf{x}_j\right) \\
&\quad - 2\sum_{i,j=1}^{n}\left(\lambda_{3i}-\lambda_{3i}^*\right)\left(\lambda_{4j}-\lambda_{4j}^*\right)K\left(\mathbf{x}_i,\mathbf{x}_j\right) \\
&\quad + \sum_{i,j=1}^{n}\left(\lambda_{3i}+\lambda_{3i}^*\right)\left(\lambda_{3j}+\lambda_{3j}^*\right)K\left(\left|\mathbf{x}_i\right|,\left|\mathbf{x}_j\right|\right) \\
&\quad + \sum_{i,j=1}^{n}\left(\lambda_{4i}+\lambda_{4i}^*\right)\left(\lambda_{4j}+\lambda_{4j}^*\right)K\left(\left|\mathbf{x}_i\right|,\left|\mathbf{x}_j\right|\right) \\
&\quad - 2\sum_{i,j=1}^{n}\left(\lambda_{3i}+\lambda_{3i}^*\right)\left(\lambda_{4j}+\lambda_{4j}^*\right)K\left(\left|\mathbf{x}_i\right|,\left|\mathbf{x}_j\right|\right) \\
&\quad + \sum_{i,j=1}^{n}\lambda_{1i}\lambda_{1j}K\left(\left|\mathbf{x}_i\right|,\left|\mathbf{x}_j\right|\right) \\
&\quad - 2\sum_{i,j=1}^{n}\lambda_{1i}\left(\lambda_{3j}+\lambda_{3j}^*\right)K\left(\left|\mathbf{x}_i\right|,\left|\mathbf{x}_j\right|\right)\Bigg) \\
& - \frac{1}{2C}\sum_{i=1}^{n}\lambda_{1i}^2 - \frac{1}{2C}\sum_{i=1}^{n}\left(\lambda_{2i}^2+\lambda_{2i}^{*2}\right) \\
& + \sum_{i=1}^{n}\left(\lambda_{2i}-\lambda_{2i}^*\right)y_i + \sum_{i=1}^{n}\left(\lambda_{3i}-\lambda_{3i}^*\right)y_i \\
& - \sum_{i=1}^{n}\left(\lambda_{4i}-\lambda_{4i}^*\right)y_i \\
& + \sum_{i=1}^{n}\left(\lambda_{3i}+\lambda_{3i}^*\right)\epsilon_i - \sum_{i=1}^{n}\left(\lambda_{4i}+\lambda_{4i}^*\right)\epsilon_i \\
\text{s.t.} \quad & \lambda_{1i},\lambda_{2i},\lambda_{2i}^*,\lambda_{3i},\lambda_{3i}^*,\lambda_{4i},\lambda_{4i}^* \ge 0,
\end{aligned}
\tag{8}
$$

where λ_{1i}, λ_{2i}, λ_{2i}^*, λ_{3i}, λ_{3i}^*, λ_{4i}, and λ_{4i}^* are Lagrange multipliers. $K(*)$ is a nonlinear kernel. The followings are well-known nonlinear kernels, where σ, γ, r, h, and θ are kernel parameters:

(1) Gaussian (radial basis) kernel: $e^{-\|x_i-x_j\|^2/2\sigma^2}$, $\sigma > 0$ [10],

(2) hyperbolic tangent kernel: $\tanh(\gamma x_i x_j^t + \theta)$, $\gamma > 0$ [12],

(3) polynomial kernel: $(\gamma x_i x_j^t + r)^h$, $h \in \mathbb{N}$, $\gamma > 0$, and $r \ge 0$ [14].

The advantage of Hong and Hwang's approach is a model-free method in the sense that there is no need to assume the underlying model function for interval nonlinear regression model with crisp inputs and interval output.

2.3. Huang's Approach. There are two problems while using the traditional SVM model. (1) Large scale: when dealing with large-scale data sets, the solution may be difficult to be found when using SVM with nonlinear kernels; (2) Unbalance: the number of samples from one class is much larger than the number of samples from the other classes. It causes the excursion of separation margin.

To resolve these problems, Huang [21] proposed a reduced support vector machine (RSVM) approach in evaluating interval regression models. RSVM has been proven more efficient than the traditional SVM in processing large-scale data.

With the principle of RSVM, the interval nonlinear regression model is listed as follows:

$$\begin{aligned}
\max\quad & -\frac{1}{2}\sum_{i,j=1}^{n}\lambda_{1i}\lambda_{1j}Q^{t}_{\cdot,K}Q_{\cdot,K} \\
& -\frac{1}{2}\sum_{i,j=1}^{n}\left(\lambda_{2i}-\lambda_{2i}^{*}\right)\left(\lambda_{2j}-\lambda_{2j}^{*}\right)K\left(\mathbf{x}_i,\mathbf{x}_j\right) \\
& -\frac{1}{2}\sum_{i,j=1}^{n}\left(\lambda_{3i}-\lambda_{3i}^{*}\right)\left(\lambda_{3j}-\lambda_{3j}^{*}\right)K\left(\mathbf{x}_i,\mathbf{x}_j\right) \\
& -\sum_{i,j=1}^{n}\lambda_{1i}Q_{\cdot,K}\left(\lambda_{2j}-\lambda_{2j}^{*}\right)\mathbf{x}_j \\
& +\sum_{i,j=1}^{n}\lambda_{1i}Q_{\cdot,K}\left(\lambda_{3j}-\lambda_{3j}^{*}\right)\mathbf{x}_j \\
& +\sum_{i,j=1}^{n}\left(\lambda_{2i}-\lambda_{2i}^{*}\right)\left(\lambda_{3j}-\lambda_{3j}^{*}\right)K\left(\mathbf{x}_i,\mathbf{x}_j\right) \\
& -\frac{1}{2}\sum_{i,j=1}^{n}\left(\lambda_{2i}+\lambda_{2i}^{*}\right)\left(\lambda_{2j}+\lambda_{2j}^{*}\right)K\left(\left|\mathbf{x}_i\right|,\left|\mathbf{x}_j\right|\right) \\
& +\sum_{i,j=1}^{n}\left(\lambda_{2i}+\lambda_{2i}^{*}\right)\left(\lambda_{3j}+\lambda_{3j}^{*}\right)K\left(\left|\mathbf{x}_i\right|,\left|\mathbf{x}_j\right|\right) \\
& -\sum_{i,j=1}^{n}\left(\lambda_{3i}+\lambda_{3i}^{*}\right)\left(\lambda_{3j}+\lambda_{3j}^{*}\right)K\left(\left|\mathbf{x}_i\right|,\left|\mathbf{x}_j\right|\right) \\
& -\frac{1}{4C}\sum_{i=1}^{n}\lambda_{1i}^{2}+\sum_{i=1}^{n}\left(\lambda_{2i}-\lambda_{2i}^{*}\right)y_i \\
& -\sum_{i=1}^{n}\left(\lambda_{3i}-\lambda_{3i}^{*}\right)y_i \\
& -\sum_{i=1}^{n}\left(\lambda_{2i}+\lambda_{2i}^{*}\right)\epsilon_i+\sum_{i=1}^{n}\left(\lambda_{3i}+\lambda_{3i}^{*}\right)\epsilon_i \\
\text{s.t.}\quad & \lambda_{1i},\lambda_{2i},\lambda_{2i}^{*},\lambda_{3i},\lambda_{3i}^{*}\geq 0,
\end{aligned} \tag{9}$$

where λ_{1i}, λ_{2i}, λ_{2i}^{*}, λ_{3i}, and λ_{3i}^{*} are Lagrange multipliers. Q is a positive semidefinite matrix in RSVM. $K(*)$ is a nonlinear kernel.

The advantage of Huang's approach is to reduce the number of support vectors by randomly selecting a subset of samples. While processing with large-scale data sets, the solution can be found easily by the proposed method with nonlinear kernels.

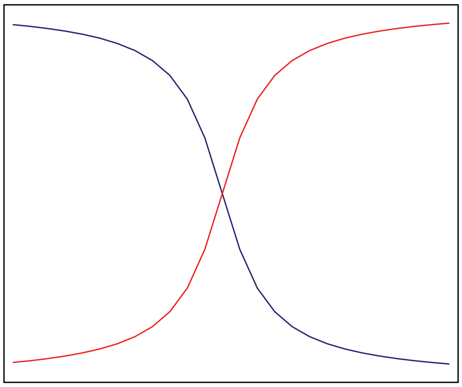

Figure 1: Soft margin.

3. Proposed Methods

In this section we first propose the soft margin method to modify the excursion of separation margin and to be effective in the gray zone. Then the formulation of interval regression with SSVM to analyze big data is introduced.

3.1. Soft Margin. In a conventional SVM, the sign function is used as the decision-making function. The separation threshold of the sign function is 0, which results in an excursion of separation margin for unbalanced data sets. The aim of the hard-margin separation margin is to find a hyperplane with the largest distance to the nearest training data. However, the limitations of the hard-margin formulation are as follows:

(1) there is no separating hyperplane for certain training data;

(2) complete separation with zero training error will lead to suboptimal prediction error;

(3) it is difficult to deal with the gray zone between classes.

Thus, the soft margin method is proposed to modify the excursion of separation margin and to be effective in the gray zone. The soft margin is defined as

$$\begin{aligned}
f^{-}(\delta) &= \frac{\arctan\left(-\delta\cdot s+\vartheta\cdot s\right)}{\pi}+0.5, \\
f^{+}(\delta) &= \frac{\arctan\left(\delta\cdot s-\vartheta\cdot s\right)}{\pi}+0.5,
\end{aligned} \tag{10}$$

where δ is the decision value. ϑ and s are offset parameter and scale parameter which need to be estimated using statistical method.

With the soft margin as shown in Figure 1, the predication of the class labels can be determined as follows:

$$\begin{aligned}
& y(x) \\
& = \begin{cases} -1, & \text{if } \left(\nu_r < f^{-}(\delta), \delta < \vartheta\right) \text{ or } \left(\nu_r > f^{+}(\delta), \delta > \vartheta\right) \\ +1, & \text{if } \left(\nu_r > f^{-}(\delta), \delta < \vartheta\right) \text{ or } \left(\nu_r < f^{+}(\delta), \delta > \vartheta\right), \end{cases}
\end{aligned} \tag{11}$$

where ν_r is a random number between 0 and 1.

3.2. Interval Regression with SSVM. The main idea of smooth support vector machine (SSVM) is solved by a fast Newton-Armijo algorithm [25] and has been extended to nonlinear separation surfaces by using a nonlinear kernel technology [24].

Suppose that m training data $\{x_i, y_i\}$, $i = 1, 2, \ldots, m$, are given, where $x_i \in \mathfrak{R}^n$ are the input patterns and $y_i \in \{-1, 1\}$ are the related target values of two-class pattern classification case. Then the standard support vector machine with a linear kernel [14] is

$$\begin{aligned} &\min_{w,b,\xi} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i^2 \\ &\text{s.t.} \quad y_i\left(w^t x_i + b\right) \geq 1 - \xi_i \\ &\qquad\quad \xi_i \geq 0, \quad i = 1, 2, \ldots, m, \end{aligned} \tag{12}$$

where b is the location of hyperplane relative to the origin. The regularization constant C is a positive parameter to control the tradeoff between the training error and the part of maximizing the margin that is achieved by minimizing $\|w\|^2$. ξ_i is the slack variable with weight $C/2$. $\|w\|$ is the Euclidean norm of w which is the normal to the following hyperplanes:

$$w^t x_i + b = +1, \quad \text{for } y_i = +1, \tag{13}$$

$$w^t x_i + b = -1, \quad \text{for } y_i = -1. \tag{14}$$

The first hyperplane (13) bounds the class $\{+1\}$ and the second hyperplane (14) bounds the class $\{-1\}$. The linear separating hyperplane is

$$w^t x_i + b = 0. \tag{15}$$

In Lee and Mangasarian's approach [24], $b^2/2$ is added to the objective function of (12). This is equivalent to adding a constant feature to the training data and finding a separating hyperplane through the origin. Consider

$$\begin{aligned} &\min_{w,b,\xi} \quad \frac{1}{2}\left(\|w\|^2 + b^2\right) + \frac{C}{2}\sum_{i=1}^{m}\xi_i^2 \\ &\text{s.t.} \quad y_i\left(w^t x_i + b\right) \geq 1 - \xi_i \\ &\qquad\quad \xi_i \geq 0, \quad i = 1, 2, \ldots, m, \end{aligned} \tag{16}$$

where $\xi_i = \{1 - y_i(w^t x_i + b)\}_+$ for all i and the "+" function is defined as $x_+ = \max\{0, x\}$. Then (12) can be reformulated as the following minimization problem by replacing ξ_i with $\{1 - y_i(w^t x_i + b)\}_+$:

$$\min_{w,b} \frac{1}{2}\left(\|w\|^2 + b^2\right) + \frac{C}{2}\sum_{i=1}^{m}\left\{1 - y_i\left(w^t x_i + b\right)\right\}_+^2. \tag{17}$$

The objective function in (17) is not twice differentiable and can be solved by using a fast Newton-Armijo method [25]. Thus the "+" function in SSVM is approximated by a smooth function, $p(x, \alpha)$, as follows:

$$p(x, \alpha) = x + \frac{1}{\alpha}\log\left(1 + e^{-\alpha x}\right), \quad \alpha > 0, \tag{18}$$

where $\alpha > 0$ is the smooth parameter. $1/(1 + e^{-\alpha x})$ is the integral of the sigmoid function of neural networks [28]. The $p(x, \alpha)$ with a smoothing parameter α is to replace the "+" function of (17) to obtain the following smooth support vector machine (SSVM) with a linear kernel:

$$\min_{w,b} \frac{1}{2}\left(\|w\|^2 + b^2\right) + \frac{C}{2}\sum_{i=1}^{m}p\left(\left\{1 - y_i\left(w^t x_i + b\right)\right\}, \alpha\right)^2. \tag{19}$$

For specific data sets, an appropriate nonlinear mapping $x \mapsto \phi(x)$ can be used to embed the original $\mathfrak{R}^n$ features into a Hilbert feature space $\mathcal{F}$, $\phi : \mathfrak{R}^n \mapsto \mathcal{F}$, with a nonlinear kernel $K(x_i, x_j) \equiv \phi(x_i)^t\phi(x_j)$. Thus, (19) can be extended to the SSVM with a nonlinear kernel:

$$\begin{aligned} \min_{w,b} \; & \frac{1}{2}\left(\|w\|^2 + b^2\right) \\ & + \frac{C}{2}\sum_{i=1}^{m}p\left(\left\{1 - y_i\left(\sum_{j=1}^{m}v_j K\left(x_i, x_j\right) + b\right)\right\}, \alpha\right)^2, \end{aligned} \tag{20}$$

where $\sum_{j=1}^{m} v_j K(x_i, x_j) + b$ is the nonlinear SSVM classifier. The coefficient v_j is determined by solving an optimization problem (20) and the data points with corresponding nonzero coefficients.

With the principle of SSVM, we can formulate the interval linear regression model as follows:

$$\begin{aligned} \min_{\overline{a},\overline{c},\overline{d}} \quad & \frac{1}{2}\left(\overline{a}^t\overline{a} + \overline{c}^t\overline{c} + \overline{d}^t\overline{d} + b^2\right) \\ & + \frac{C}{2}\sum_{i=1}^{m}p\left(\left\{1 - y_i\left(w^t x_i + b\right)\right\}, \alpha\right)^2 \\ \text{s.t.} \quad & \overline{a}\mathbf{x}_i + \overline{c}\left|\mathbf{x}_i\right| \leq y_i + e_i \\ & \overline{a}\mathbf{x}_i - \overline{c}\left|\mathbf{x}_i\right| \geq y_i - e_i \\ & \overline{a}\mathbf{x}_i + \overline{c}\left|\mathbf{x}_i\right| + \overline{d}\left|\mathbf{x}_i\right| \geq y_i + e_i \\ & \overline{a}\mathbf{x}_i - \overline{c}\left|\mathbf{x}_i\right| - \overline{d}\left|\mathbf{x}_i\right| \leq y_i - e_i \\ & i = 1, 2, \ldots, m, \end{aligned} \tag{21}$$

where $\overline{a}$, $\overline{c}$, and $\overline{d}$ are the collections of all a_i, c_i, and d_i, $i = 1, 2, \ldots, m$, respectively.

Given (21), the corresponding Lagrangian objective function is

$$\begin{aligned} L := {} & \frac{1}{2}\left(\bar{a}^t\bar{a} + \bar{c}^t\bar{c} + \bar{d}^t\bar{d} + b^2\right) \\ & + \frac{C}{2}\sum_{i=1}^{m} p\left(\left\{1 - y_i\left(w^t x_i + b\right)\right\}, \alpha\right)^2 \\ & - \sum_{i=1}^{m}\lambda_{1i}\left(y_i + e_i - \bar{a}\mathbf{x}_i - \bar{c}\left|\mathbf{x}_i\right|\right) \\ & - \sum_{i=1}^{m}\lambda_{2i}\left(\bar{a}\mathbf{x}_i - \bar{c}\left|\mathbf{x}_i\right| - y_i + e_i\right) \\ & - \sum_{i=1}^{m}\lambda_{3i}\left(\bar{a}\mathbf{x}_i + \bar{c}\left|\mathbf{x}_i\right| + \bar{d}\left|\mathbf{x}_i\right| - y_i - e_i\right) \\ & - \sum_{i=1}^{m}\lambda_{4i}\left(y_i - e_i - \bar{a}\mathbf{x}_i + \bar{c}\left|\mathbf{x}_i\right| + \bar{d}\left|\mathbf{x}_i\right|\right), \end{aligned} \tag{22}$$

where L is Lagrangian and λ_{1i}, λ_{2i}, λ_{3i}, and λ_{4i} are Lagrange multipliers. The idea to construct a Lagrange function from the objective function and the corresponding constraints is to introduce a dual set of variables. It can be shown that the Lagrangian function has a saddle point with respect to the primal and dual variables in the solution [29].

The Karush-Kuhn-Tucker (KKT) conditions that the partial derivatives of L with respect to the primal variables $(\bar{a}, \bar{c}, \bar{d})$ for optimality

$$\begin{aligned} \frac{\partial L}{\partial \bar{a}} &= 0 \Longrightarrow \bar{a} = -\sum_{i=1}^{m}\left(\lambda_{1i} - \lambda_{2i} - \lambda_{3i} + \lambda_{4i}\right)\mathbf{x}_i, \\ \frac{\partial L}{\partial \bar{c}} &= 0 \Longrightarrow \bar{c} = -\sum_{i=1}^{m}\left(\lambda_{1i} + \lambda_{2i} - \lambda_{3i} - \lambda_{4i}\right)\left|\mathbf{x}_i\right|, \\ \frac{\partial L}{\partial \bar{d}} &= 0 \Longrightarrow \bar{d} = \sum_{i=1}^{m}\left(\lambda_{3i} + \lambda_{4i}\right)\left|\mathbf{x}_i\right|. \end{aligned} \tag{23}$$

Substituting (23) in (22) yields the following optimization problem:

$$\begin{aligned} \max \quad & \frac{1}{2}\left(\sum_{i,j=1}^{m}\left(\lambda_{1i} - \lambda_{2i} - \lambda_{3i} + \lambda_{4i}\right)\right. \\ & \qquad \times\left(\lambda_{1j} - \lambda_{2j} - \lambda_{3j} + \lambda_{4j}\right)\mathbf{x}_i^t\mathbf{x}_j \\ & \quad + \sum_{i,j=1}^{m}\left(\lambda_{1i} + \lambda_{2i} - \lambda_{3i} - \lambda_{4i}\right) \\ & \qquad \times\left(\lambda_{1j} + \lambda_{2j} - \lambda_{3j} - \lambda_{4j}\right)\left|\mathbf{x}_i\right|^t\left|\mathbf{x}_j\right| \\ & \quad \left. - \sum_{i,j=1}^{m}\left(\lambda_{3i} + \lambda_{4i}\right)\left(\lambda_{3j} + \lambda_{4j}\right)\left|\mathbf{x}_i\right|^t\left|\mathbf{x}_j\right| + b^2\right) \\ & + \frac{C}{2}\sum_{i=1}^{m} p\left(\left\{1 - y_i\left(w^t x_i + b\right)\right\}, \alpha\right)^2 \\ \text{s.t.} \quad & \lambda_{1i}, \lambda_{2i}, \lambda_{3i}, \lambda_{4i} \geq 0. \end{aligned} \tag{24}$$

Similarly, we can obtain the interval nonlinear regression model by mapping $x \mapsto \phi(x)$ to embed the original $\mathfrak{R}^n$ features into a Hilbert feature space $\mathcal{F}$, $\phi : \mathfrak{R}^n \mapsto \mathcal{F}$, with a nonlinear kernel $K(x_i, x_j) \equiv \phi(x_i)^t\phi(x_j)$ as discussed in Section 2.2. Then we obtain the optimization problem as (25) by replacing $\mathbf{x}_i^t\mathbf{x}_j$ and $|\mathbf{x}_i|^t|\mathbf{x}_j|$ in (24) with $K(\mathbf{x}_i, \mathbf{x}_j)$ and $K(|\mathbf{x}_i|, |\mathbf{x}_j|)$, respectively:

$$\begin{aligned} \max \quad & \frac{1}{2}\left(\sum_{i,j=1}^{m}\left(\lambda_{1i} - \lambda_{2i} - \lambda_{3i} + \lambda_{4i}\right)\right. \\ & \qquad \times\left(\lambda_{1j} - \lambda_{2j} - \lambda_{3j} + \lambda_{4j}\right)K\left(\mathbf{x}_i, \mathbf{x}_j\right) \\ & \quad + \sum_{i,j=1}^{m}\left(\lambda_{1i} + \lambda_{2i} - \lambda_{3i} - \lambda_{4i}\right) \\ & \qquad \times\left(\lambda_{1j} + \lambda_{2j} - \lambda_{3j} - \lambda_{4j}\right)K\left(\left|\mathbf{x}_i\right|, \left|\mathbf{x}_j\right|\right) \\ & \quad - \sum_{i,j=1}^{m}\left(\lambda_{3i} + \lambda_{4i}\right) \\ & \qquad \left.\times\left(\lambda_{3j} + \lambda_{4j}\right)K\left(\left|\mathbf{x}_i\right|, \left|\mathbf{x}_j\right|\right) + b^2\right) \\ & + \frac{C}{2}\sum_{i=1}^{m} p\left(\left\{1 - y_i\left(\sum_{j=1}^{m} v_j K\left(\mathbf{x}_i, \mathbf{x}_j\right) + b\right)\right\}, \alpha\right)^2 \\ \text{s.t.} \quad & \lambda_{1i}, \lambda_{2i}, \lambda_{3i}, \lambda_{4i} \geq 0. \end{aligned} \tag{25}$$

4. Numerical Example

To illustrate the methods developed in Section 3, the following example is presented.

Example. To illustrate the proposed methods dealing with big data sets, we use the data sets from Taiwan Stock Exchange Capitalization Weighted Stock Index (TAIEX) [26] which included the highest, lowest, and closed data and the ranges are from 01/02/2012 to 12/28/2012, from 01/02/2011 to 12/28/2012, from 01/02/2010 to 12/28/2012, and from 01/02/2009 to 12/28/2012, respectively. For these data sets, the Gaussian kernel [10] is used where $\sigma = 2.5$ and the regularization constant $C = 300$. The results are illustrated from Figure 2 to Figure 5.

The comparison is shown by using the measure of fitness [15] as (26), which defines how closely the possibility output

Table 1: Comparison results of the measure of fitness.

	Tanaka and Lee [15]	Hong and Hwang [16]	Huang [21]	Proposed methods
φ_Y (Figure 2)	0.1404	0.1395	0.1412	0.1354
φ_Y (Figure 3)	0.1573	0.1562	0.1581	0.1429
φ_Y (Figure 4)	0.1694	0.1658	0.1706	0.1583
φ_Y (Figure 5)	0.1714	0.1695	0.1723	0.1609

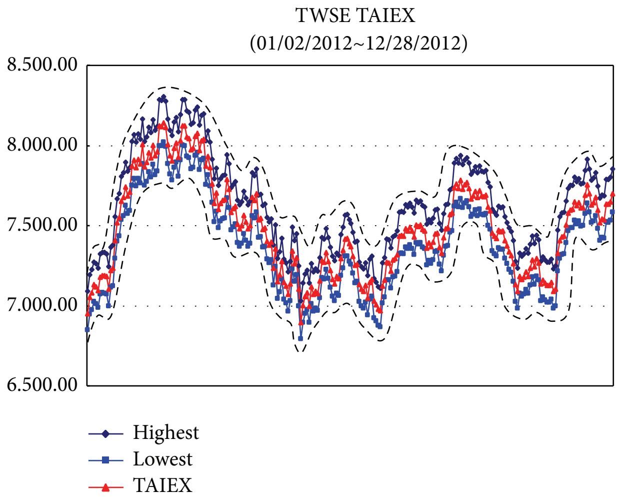

Figure 2: TAIEX [26] from 01/02/2012 to 12/28/2012.

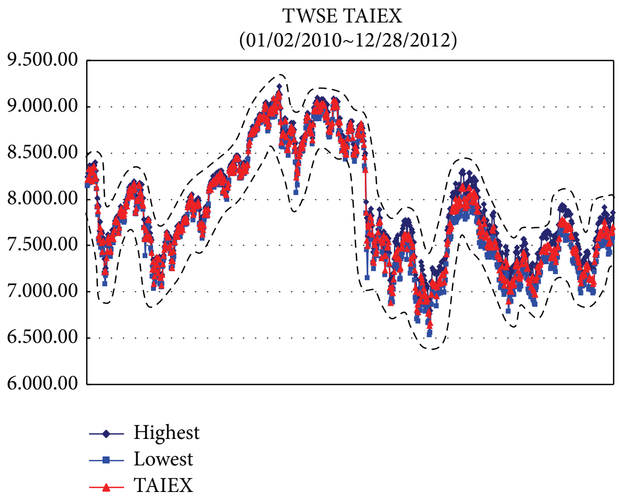

Figure 4: TAIEX [26] from 01/02/2010 to 12/28/2012.

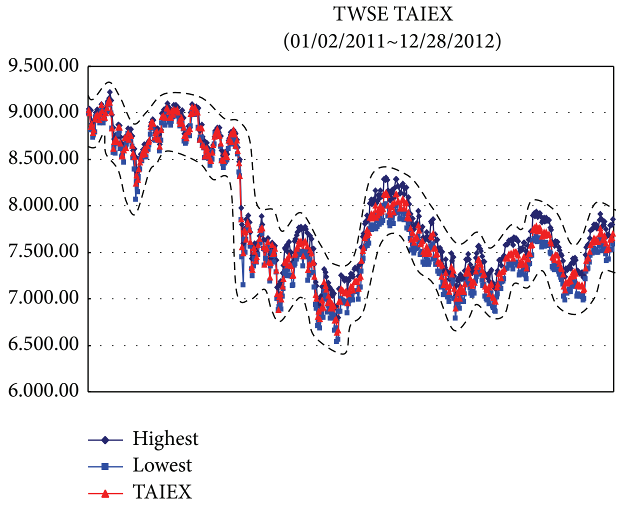

Figure 3: TAIEX [26] from 01/02/2011 to 12/28/2012.

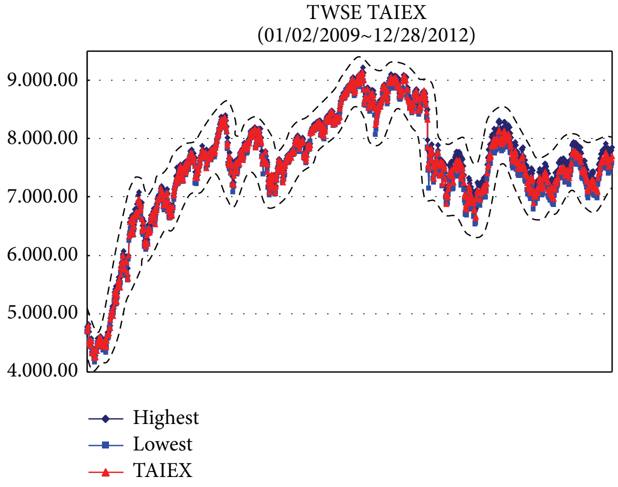

Figure 5: TAIEX [26] from 01/02/2009 to 12/28/2012.

for the jth input approximates the necessity output for the jth input. Consider

$$\varphi_Y(\mathbf{x}_i) = \frac{1}{q}\sum_{j=1}^{q}\frac{c_0+\sum_{i=1}^{n} c_i\left|x_{ij}\right|}{c_0+\sum_{i=1}^{n} c_i\left|x_{ij}\right|+d_0+\sum_{i=1}^{n} d_i\left|x_{ij}\right|}, \qquad (26)$$

where q is a sample size and $0 \le \varphi_Y \le 1$.

Table 1 presents the proposed methods with a Gaussian kernel along with the results computed by Tanaka and Lee [15], Hong and Hwang [16], and Huang [21]. We can find that the proposed methods are more efficient than other methods.

5. Conclusions

In this paper, we collaborate interval regression with SSVM to analyze big data. In addition, the soft margin method is proposed to modify the excursion of separation margin and to be effective in the gray zone. The main idea of SSVM is solved by a fast Newton-Armijo algorithm and has been extended to nonlinear separation surfaces by using a nonlinear kernel technology. The experiment results show that the proposed methods are more efficient than existing methods. In this study, we estimate the interval regression model with crisp inputs and interval output. In future works, both interval inputs-interval output and fuzzy inputs-fuzzy output will be considered.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors appreciate the anonymous referees for their useful comments and suggestions which helped to improve the quality and presentation of this paper. The original version was accepted by International Conference on Business, Information, and Cultural Creative Industry, 2014. Also, special thanks are due to the National Science Council, Taiwan, for financially supporting this research under Grants nos. NSC 102-2410-H-141-012-MY2 (C. H. Huang) and NSC 102-2410-H-259-039-(H. Y. Kao).

References

[1] D. Laney, *The Importance of Big Data: A Definition*, Gartner, 2012.

[2] C. L. P. Chen and C. Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," *Information Sciences*, vol. 275, pp. 314–347, 2014.

[3] K. Kambatla, G. Kollias, V. Kumar, and A. Grama, "Trends in big data analytics," *Journal of Parallel and Distributed Computing*, vol. 74, no. 7, pp. 2561–2573, 2014.

[4] V. López, S. del Ro, J. M. Bentez, and F. Herrera, "Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced big data," *Fuzzy Sets and Systems*, 2014.

[5] T. Shelton, A. Poorthuis, M. Graham, and M. Zook, "Mapping the data shadows of Hurricane Sandy: uncovering the sociospatial dimensions of 'big data,'" *Geoforum*, vol. 52, pp. 167–179, 2014.

[6] M. Arun Kumar, R. Khemchandani, M. Gopal, and S. Chandra, "Knowledge based least squares twin support vector machines," *Information Sciences*, vol. 180, no. 23, pp. 4606–4618, 2010.

[7] S. Maldonado, R. Weber, and J. Basak, "Simultaneous feature selection and classification using kernel-penalized support vector machines," *Information Sciences*, vol. 181, no. 1, pp. 115–128, 2011.

[8] O. L. Mangasarian, "Mathematical programming in data mining," *Data Mining and Knowledge Discovery*, vol. 1, no. 2, pp. 183–201, 1997.

[9] O. L. Mangasarian, "Generalized support vector machines," in *Advances in Large Margin Classifiers*, A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, Eds., pp. 135–146, The MIT Press, Cambridge, Mass, USA, 2000.

[10] C. A. Micchelli, "Interpolation of scattered data: distance matrices and conditionally positive definite functions," *Constructive Approximation*, vol. 2, no. 1, pp. 11–22, 1986.

[11] R. Savitha, S. Suresh, and N. Sundararajan, "Fast learning Circular COMplex-valued Extreme Learning Machine (CCELM) for real-valued classification problems," *Information Sciences*, vol. 187, pp. 277–290, 2012.

[12] B. Schölkopf, C. J. C. Burges, and A. J. Smola, *Advances in Kernel Methods: Support Vector Learning*, MIT Press, Cambridge, Mass, USA, 1999.

[13] A. Unler, A. Murat, and R. B. Chinnam, "mr^{2}PSO: a maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification," *Information Sciences*, vol. 181, no. 20, pp. 4625–4641, 2011.

[14] V. N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, NY, USA, 1998.

[15] H. Tanaka and H. Lee, "Interval regression analysis by quadratic programming approach," *IEEE Transactions on Fuzzy Systems*, vol. 6, no. 4, pp. 473–481, 1998.

[16] D. H. Hong and C. H. Hwang, "Interval regression analysis using quadratic loss support vector machine," *IEEE Transactions on Fuzzy Systems*, vol. 13, no. 2, pp. 229–237, 2005.

[17] A. Bisserier, R. Boukezzoula, and S. Galichet, "A revisited approach to linear fuzzy regression using trapezoidal fuzzy intervals," *Information Sciences*, vol. 180, no. 19, pp. 3653–3673, 2010.

[18] P. D'Urso, R. Massari, and A. Santoro, "Robust fuzzy regression analysis," *Information Sciences*, vol. 181, no. 19, pp. 4154–4174, 2011.

[19] J. T. Jeng, C. C. Chuang, and S. F. Su, "Support vector interval regression networks for interval regression analysis," *Fuzzy Sets and Systems*, vol. 138, no. 2, pp. 283–300, 2003.

[20] C. Huang and H. Kao, "Interval regression analysis with soft-margin reduced support vector machine," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5579, pp. 826–835, 2009.

[21] C. H. Huang, "A reduced support vector machine approach for interval regression analysis," *Information Sciences*, vol. 217, pp. 56–64, 2012.

[22] C.-C. Chang, L.-J. Chien, and Y.-J. Lee, "A novel framework for multi-class classification via ternary smooth support vector machine," *Pattern Recognition*, vol. 44, no. 6, pp. 1235–1244, 2011.

[23] Y. J. Lee, W. F. Hsieh, and C. M. Huang, "ε-SSVR: a smooth support vector machine for ε-insensitive regression," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 5, pp. 678–685, 2005.

[24] Y. Lee and O. L. Mangasarian, "SSVM: a smooth support vector machine for classification," *Computational Optimization and Applications*, vol. 20, no. 1, pp. 5–22, 2001.

[25] L. Armijo, "Minimization of functions having Lipschitz continuous first partial derivatives," *Pacific Journal of Mathematics*, vol. 16, pp. 1–3, 1966.

[26] "Taiwan Stock Exchange Capitalization Weighted Stock Index," http://www.twse.com.tw.

[27] H. Tanaka, S. Uejima, and K. Asai, "Fuzzy linear regression model," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 10, pp. 2933–2938, 1980.

[28] O. L. Mangasarian, "Mathematical programming in neural networks," *ORSA Journal on Computing*, vol. 5, no. 4, pp. 349–360, 1993.

[29] O. L. Mangasarian, *Nonlinear Programming*, McGraw-Hill, New York, NY, USA, 1969.