Support Vector Machine for Interval Regression

Dug Hun Hong[1], Changha Hwang[2]

Abstract

Support vector machine (SVM) has been very successful in pattern recognition and function estimation problems for crisp data. This paper proposes a new method to evaluate interval linear and nonlinear regression models combining the possibility and necessity estimation formulation with the principle of SVM. For data sets with crisp inputs and interval outputs, the possibility and necessity models have been recently utilized, which are based on quadratic programming approach giving more diverse spread coefficients than a linear programming one. SVM also uses quadratic programming approach whose another advantage in interval regression analysis is to be able to integrate both the property of central tendency in least squares and the possibilistic property in fuzzy regression. However this is not a computationally expensive way. SVM allows us to perform interval nonlinear regression analysis by constructing an interval linear regression function in a high dimensional feature space. In particular, SVM is a very attractive approach to model nonlinear interval data. The proposed algorithm here is model-free method in the sense that we do not have to assume the underlying model function for interval nonlinear regression model with crisp inputs and interval output. Experimental results are then presented which indicate the performance of this algorithm.

Key words : Interval Regression Analysis, Possibility and Necessity Models, Quadratic Programming, Support Verctor Machine.

1. Introduction

In many real applications, information is often uncertain, imprecise and incomplete, which can be represented by fuzzy data or a generalization of interval data. For handling interval data, fuzzy regression analysis becomes an important tool and successfully applied to different applications such as market forecasting, system identification, and so on. Fuzzy regression analysis can be simplified to interval regression analysis, where interval regression models are implemented. In interval linear regression, possibility and necessity models have been employed under given interval data. Coefficients in interval regression model are assumed to be interval. In fact, interval regression is regarded as the simplest version of possibilistic regression analysis. Possibilistic regression analysis has been first proposed by Tanaka et al.(1992) where a fuzzy linear system has been used as a regression

1) Dept. of Mathematics, Myungji University
2) Dept. of Statistical Information, Catholic University of Daehu

model. To determine the interval parameters of interval regression models, a basic linear programming (LP) or quadratic programming (QP) problem should be solved.

In previous LP-based approaches, some coefficients become crisp because of the characteristic of LP, since these regression analyses have been reduced to LP problems. To overcome this crisp characteristic of LP, Tanaka and Lee(1998) propose interval regression analysis based on QP. They adopt the dual method for solving their QP formulations. First, they introduce a basic formulation where the sum of squared residuals of the estimated interval outputs is considered as an objective function. This QP based approach is more flexible than the LP based one in the sense that the noncrisp coefficients obtained by QP are more desirable than those by LP. Next, they introduce another QP formulation that combines the property of central tendency in least squares and the possibilistic property in fuzzy regression. This approach has more central tendency than their previous approaches.

In this paper, we propose a new method to evaluate interval linear and nonlinear regression models combining the possibility and necessity estimation formulation with the principle of support vector machine (SVM). This proposed SVM is also based on QP approach giving fairly diverse spread coefficients and integrates well both the property of central tendency in least squares and the possibilistic property in fuzzy regression. However, this is not a computationally expensive way. This SVM allows us to perform interval nonlinear regression analysis by constructing an interval linear regression function in a high dimensional feature space. In particular, SVM is a very attractive approach to model nonlinear interval data. The proposed algorithm here is model-free method in the sense that we do not have to assume the underlying model function for interval nonlinear regression model with crisp inputs and interval output. This model-free method turns out to be a promissing method which has been attempted to treat interval nonlinear regression model.

2. Interval Regression for Crisp Input-Interval Output

In this section, we need to briefly look at how to get solutions for interval regression models using QP approach proposed by Tanaka and Lee(1998). For a data set with crisp inputs and interval outputs, we can consider two interval regression models, i.e., the possibility and necessity models. In this section, we review the unified QP approach to obtain the possibility and necessity models simultaneously. In this unified approach, we assume for simplicity that the center coefficients of the possibility regression model and necessity regression model are same.

Suppose that we are given training data $\{(\boldsymbol{x}_i, Y_i),\ i=1,\cdots,n\}$, where $\boldsymbol{x}_i = (1, x_{i1},\cdots,x_{im})^t$ is the ith input vector, $Y_i = (y_i, e_i)$ is the corresponding interval output that consists of a center y_i and a radius e_i. For this data set, the possibility and necessity estimation models are denoted as

$$Y^*(\boldsymbol{x}_i) = A_0^* + A_1^* x_{i1} + \cdots + A_m^* x_{im},\ i=1,\cdots,n$$

$$Y_*(\boldsymbol{x}_i) = A_{0*} + A_{1*} x_{i1} + \cdots + A_{*m} x_{im},\ i=1,\cdots,n$$

where the interval coefficients A_i^* and A_{*i} are denoted as $A_i^* = (a_i^*, c_i^*)$ and $A_{*i} = (a_{*i}, c_{*i})$, respectively. The estimated interval $Y^*(\boldsymbol{x}_i)$ by the possibility model always includes the observed interval Y_i, whereas the estimated interval $Y_*(\boldsymbol{x}_i)$ by the

necessity model should be included in the observed interval Y_i. In fact, we can denote the interval coefficients A_i^* and A_{*i} as

$$A_i^* = (a_i, c_i + d_i),\ A_{*i} = (a_i, c_i)$$

which satisfies the condition $A_{*i} \subseteq A_i^*, i = 0,1,\cdots,m$ since c_i and d_i are assumed to be positive. Therefore, by interval arithmetic the possibility model $Y^*(\boldsymbol{x}_i)$ and the necessity model $Y_*(\boldsymbol{x}_i)$ can be written as

$$Y^*(\boldsymbol{x}_i) = (\boldsymbol{a}^t\boldsymbol{x}_i, \boldsymbol{c}^t|\boldsymbol{x}_i| + \boldsymbol{d}^t|\boldsymbol{x}_i|),\quad Y_*(\boldsymbol{x}_i) = (\boldsymbol{a}^t\boldsymbol{x}_i, \boldsymbol{c}^t|\boldsymbol{x}_i|).$$

Then, the objective function in the unified approach by QP can be assumed as the following quadratic function:

$$min_{a,c,d}\ J = \boldsymbol{d}^t\left(\sum_{i=1}^{n} |\boldsymbol{x}_i|\boldsymbol{x}_i|^t\right)\boldsymbol{d} + \xi(\boldsymbol{a}^t b + \boldsymbol{c}^t \boldsymbol{c}).$$

3. Quadratic Loss SVM for Interval Regression

In this section, we propose a new method to evaluate interval linear and nonlinear regression models combining the possibility estimation formulation integrating the property of central tendency with the principle of SVM. We first need to look at how to get solutions for interval linear regression models by implementing the SVM approach. We follow the way of constructing objective function in SVM regression. Then, the objective function can be assumed as the following quadratic function:

$$\min_{a,c}\ \frac{1}{2}(\|\boldsymbol{a}\|^2 + \|\boldsymbol{c}\|^2 + \|\boldsymbol{d}\|^2) + \frac{\gamma}{2}\left(\sum_{i=1}^{n}\xi_{1i}^2 + \sum_{i=1}^{n}(\xi_{2i}^2 + \xi_{2i}^{*2})\right) \quad (1)$$

subject to

$$\boldsymbol{d}^t|\boldsymbol{x}_i| \le \xi_{1i}$$

$$y_i - \boldsymbol{a}^t\boldsymbol{x}_i \le \xi_{2i},\quad \boldsymbol{a}^t\boldsymbol{x}_i - y_i \le \xi_{2i}^*,\ i = 1,\cdots,n$$

$$\boldsymbol{a}^t\boldsymbol{x}_i + \boldsymbol{c}^t|\boldsymbol{x}_i| + \boldsymbol{d}^t|\boldsymbol{x}_i| \ge y_i + e_i,\quad \boldsymbol{a}^t\boldsymbol{x}_i - \boldsymbol{c}^t|\boldsymbol{x}_i| - \boldsymbol{d}^t|\boldsymbol{x}_i| \le y_i - e_i,\ i = 1,\cdots,n.$$

$$\boldsymbol{a}^t\boldsymbol{x}_i + \boldsymbol{c}^t|\boldsymbol{x}_i| \le y_i + e_i,\quad \boldsymbol{a}^t\boldsymbol{x}_i - \boldsymbol{c}^t|\boldsymbol{x}_i| \ge y_i - e_i,\ i = 1,\cdots,n.$$

Although it is possible to use two weight coefficients like Tanaka and Lee(1998), we use one weight coefficient. Here, ξ_{1i} represents spreads of the estimated outputs, and ξ_{2i}, ξ_{2i}^* are slack variables representing upper and lower constraints on the outputs of the model. Hence, we can construct a Lagrange function as follows:

$$\begin{aligned} L = {} & \frac{1}{2}(\|\boldsymbol{a}\|^2 + \|\boldsymbol{c}\|^2 + \|\boldsymbol{d}\|^2) + \frac{\gamma}{2}\left(\sum_{i=1}^{n}\xi_{1i}^2 + \sum_{i=1}^{n}(\xi_{2i}^2 + \xi_{2i}^{*2})\right) \\ & - \sum_{i=1}^{n}\alpha_{1i}(\xi_{1i} - \boldsymbol{d}^t|\boldsymbol{x}_i|) \qquad (2) \\ & - \sum_{i=1}^{n}\alpha_{2i}(\xi_{2i} - y_i + \boldsymbol{a}^t\boldsymbol{x}_i) - \sum_{i=1}^{n}\alpha_{2i}^*(\xi_{2i} - \boldsymbol{a}^t\boldsymbol{x}_i + y_i) \\ & - \sum_{i=1}^{n}\alpha_{3i}(\boldsymbol{a}^t\boldsymbol{x}_i + \boldsymbol{c}^t|\boldsymbol{x}_i| + \boldsymbol{d}^t|\boldsymbol{x}_i| - y_i - e_i) - \sum_{i=1}^{n}\alpha_{3i}^*(y_i - e_i - \boldsymbol{a}^t\boldsymbol{x}_i + \boldsymbol{c}^t|\boldsymbol{x}_i| + \boldsymbol{d}^t|\boldsymbol{x}_i|) \\ & - \sum_{i=1}^{n}\alpha_{4i}(y_i + e_i - \boldsymbol{a}^t\boldsymbol{x}_i - \boldsymbol{c}^t|\boldsymbol{x}_i|) - \sum_{i=1}^{n}\alpha_{4i}^*(\boldsymbol{a}^t\boldsymbol{x}_i - \boldsymbol{c}^t|\boldsymbol{x}_i| - y_i + e_i) \end{aligned}$$

Here, $\alpha_{1i}, \alpha_{2i}, \alpha_{2i}^*, \alpha_{3i}, \alpha_{3i}^*, \alpha_{4i}, \alpha_{4i}^*$ are Lagrange multipliers. It follows from the saddle point condition that the partial derivatives of L with respect to the primal variables

$(\boldsymbol{a}, c, \boldsymbol{d}, \xi_{1i}, \xi_{2i}, \xi^*_{2i})$ have to vanish for optimality.

$$\frac{\partial L}{\partial \boldsymbol{a}} = 0 \rightarrow \boldsymbol{a} = \sum_{i=1}^{n}(\alpha_{2i} - \alpha^*_{2i})\boldsymbol{x}_i + \sum_{i=1}^{n}(\alpha_{3i} - \alpha^*_{3i})\boldsymbol{x}_i - \sum_{i=1}^{n}(\alpha_{4i} - \alpha^*_{4i})\boldsymbol{x}_i \tag{3}$$

$$\frac{\partial L}{\partial c} = 0 \rightarrow c = \sum_{i=1}^{n}(\alpha_{3i} + \alpha^*_{3i})|\boldsymbol{x}_i| - \sum_{i=1}^{n}(\alpha_{4i} + \alpha^*_{4i})|\boldsymbol{x}_i| \tag{4}$$

$$\frac{\partial L}{\partial \boldsymbol{d}} = 0 \rightarrow c = -\sum_{i=1}^{n}\alpha_{1i}|\boldsymbol{x}_i| + \sum_{i=1}^{n}(\alpha_{3i} + \alpha^*_{3i})|\boldsymbol{x}_i| \tag{5}$$

$$\frac{\partial L}{\partial \xi_{1i}} = 0 \rightarrow \xi_{1i} = \frac{1}{\gamma}\alpha_{1i} \tag{6}$$

$$\frac{\partial L}{\partial \xi^{(*)}_{2i}} = 0 \rightarrow \xi^{(*)}_{2i} = \frac{1}{\gamma}\alpha^{(*)}_{2i} \tag{7}$$

Substituting (3)-(7) into (2) yields the dual optimization problem.

$$\begin{aligned}
\text{maximize}\Big\{ & -\frac{1}{2}\Big(\sum_{i,j=1}^{n}(\alpha_{2i} - \alpha^*_{2i})(\alpha_{2j} - \alpha^*_{2j})\boldsymbol{x}_i^t\boldsymbol{x}_j \\
& + \sum_{i,j=1}^{n}(\alpha_{3i} - \alpha^*_{3i})(\alpha_{3j} - \alpha^*_{3j})\boldsymbol{x}_i^t\boldsymbol{x}_j + \sum_{i,j=1}^{n}(\alpha_{4i} - \alpha^*_{4i})(\alpha_{4j} - \alpha^*_{4j})\boldsymbol{x}_i^t\boldsymbol{x}_j \\
& + 2\sum_{i,j=1}^{n}(\alpha_{2i} - \alpha^*_{2i})(\alpha_{3j} - \alpha^*_{3j})\boldsymbol{x}_i^t\boldsymbol{x}_j - 2\sum_{i,j=1}^{n}(\alpha_{2i} - \alpha^*_{2i})(\alpha_{4j} - \alpha^*_{4j})\boldsymbol{x}_i^t\boldsymbol{x}_j \\
& - 2\sum_{i,j=1}^{n}(\alpha_{3i} - \alpha^*_{3i})(\alpha_{4j} - \alpha^*_{4j})\boldsymbol{x}_i^t\boldsymbol{x}_j + \sum_{i,j=1}^{n}(\alpha_{3i} + \alpha^*_{3i})(\alpha_{3j} + \alpha^*_{3j})|\boldsymbol{x}_i|^t|\boldsymbol{x}_j| \\
& + \sum_{i,j=1}^{n}(\alpha_{4i} + \alpha^*_{4i})(\alpha_{4j} + \alpha^*_{4j})|\boldsymbol{x}_i|^t|\boldsymbol{x}_j| - 2\sum_{i,j=1}^{n}(\alpha_{3i} + \alpha^*_{3i})(\alpha_{4j} + \alpha^*_{4j})|\boldsymbol{x}_i|^t|\boldsymbol{x}_j| \\
& + \sum_{i,j=1}^{n}\alpha_{1i}\alpha_{1j}|\boldsymbol{x}_i|^t|\boldsymbol{x}_j| - 2\sum_{i,j=1}^{n}\alpha_{1i}(\alpha_{3j} + \alpha^*_{3j})|\boldsymbol{x}_i|^t|\boldsymbol{x}_j|\Big) \\
& + \sum_{i,j=1}^{n}(\alpha_{3i} + \alpha^*_{3i})(\alpha_{3j} + \alpha^*_{3j})|\boldsymbol{x}_i|^t|\boldsymbol{x}_j|\Big) - \frac{1}{2C}\sum_{i=1}^{n}\alpha^2_{1i} \\
& - \frac{1}{2C}\sum_{i=1}^{n}\alpha^2_{1i} - \frac{1}{2C}\sum_{i=1}^{n}(\alpha^2_{2i} + \alpha^{*2}_{2i}) \\
& + \sum_{i=1}^{n}(\alpha_{2i} - \alpha^*_{2i})y_i + \sum_{i=1}^{n}(\alpha_{3i} - \alpha^*_{3i})y_i - \sum_{i=1}^{n}(\alpha_{4i} - \alpha^*_{4i})y_i \\
& + \sum_{i=1}^{n}(\alpha_{3i} + \alpha^*_{3i})e_i - \sum_{i=1}^{n}(\alpha_{4i} + \alpha^*_{4i})e_i\Big\}
\end{aligned} \tag{8}$$

subject to

$$\alpha_{1i}, \alpha_{ki}, \alpha^*_{ki} \geq 0, \; k = 2,3,4.$$

Solving (8) with above constraints determines the Lagrange multipliers, α_{1i}, $\alpha_{ki}, \alpha^*_{ki}$. Hence, if $\boldsymbol{c}^t|\boldsymbol{x}| \geq 0$ and $\boldsymbol{d}^t|\boldsymbol{x}| \geq 0$, then the linear interval regression function is as follows:

$$Y^*(\boldsymbol{x}) = (\boldsymbol{a}^t\boldsymbol{x}, \; \boldsymbol{c}^t|\boldsymbol{x}| + \boldsymbol{d}^t|\boldsymbol{x}|) \tag{9}$$

$$Y_*(\boldsymbol{x}) = (\boldsymbol{a}^t\boldsymbol{x}, \; \boldsymbol{c}^t|\boldsymbol{x}|) \tag{10}$$

Next, we will consider nonlinear interval regression model. In contrast to linear interval regression, there have been no articles on nonlinear interval regression. In this paper we treat nonlinear interval regression, without assuming the underlying model function. In the case where a linear regression function is inappropriate SVM makes algorithm nonlinear. This could be achieved by simply preprocessing input patterns $\boldsymbol{x}_i$ by a map $\Phi : R^d \rightarrow E$ into some feature space E and then applying SVM regression algorithm. This is an astonishingly straightforward way.

First notice that the only way in which the data appears in (8) is in the form of inner products $\boldsymbol{x}_i^t\boldsymbol{x}_j$, $|\boldsymbol{x}_i|^t|\boldsymbol{x}_j|$. The algorithm would only depend on the data through dot

products in E, i.e. on functions of the form $K(x_i, x_j) = .\Phi(x_i)^t\Phi(x_j)$, $K(|x_i|, |x_j|) = \Phi(|x_i|)^t\Phi(|x_j|)$. The well used kernels for regression problem are given below.

$$K(x, y) = (x^t y + 1)^p, \quad K(x, y) = e^{-\frac{|x-y|^2}{2\sigma^2}}.$$

Here, p and σ^2 are kernel parameters. In final, the nonlinear interval regression solution is given by

$$\begin{aligned}
&\text{maximize}\Big\{-\frac{1}{2}\Big(\sum_{i,j=1}^{n}(\alpha_{2i}-\alpha_{2i}^*)(\alpha_{2j}-\alpha_{2j}^*)K(x_i, x_j)\\
&+\sum_{i,j=1}^{n}(\alpha_{3i}-\alpha_{3i}^*)(\alpha_{3j}-\alpha_{3j}^*)K(x_i, x_j) + \sum_{i,j=1}^{n}(\alpha_{4i}-\alpha_{4i}^*)(\alpha_{4j}-\alpha_{4j}^*)K(x_i, x_j)\\
&+2\sum_{i,j=1}^{n}(\alpha_{2i}-\alpha_{2i}^*)(\alpha_{3j}-\alpha_{3j}^*)K(x_i, x_j) - 2\sum_{i,j=1}^{n}(\alpha_{2i}-\alpha_{2i}^*)(\alpha_{4j}-\alpha_{4j}^*)K(x_i, x_j)\\
&-2\sum_{i,j=1}^{n}(\alpha_{3i}-\alpha_{3i}^*)(\alpha_{4j}-\alpha_{4j}^*)K(x_i, x_j) + \sum_{i,j=1}^{n}(\alpha_{3i}+\alpha_{3i}^*)(\alpha_{3j}+\alpha_{3j}^*)K(|x_i|, |x_j|)\\
&+\sum_{i,j=1}^{n}(\alpha_{4i}+\alpha_{4i}^*)(\alpha_{4j}+\alpha_{4j}^*)K(|x_i|, |x_j|) - 2\sum_{i,j=1}^{n}(\alpha_{3i}+\alpha_{3i}^*)(\alpha_{4j}+\alpha_{4j}^*)K(|x_i|, |x_j|)\\
&+\sum_{i,j=1}^{n}\alpha_{1i}\alpha_{1j}K(|x_i|, |x_j|) - 2\sum_{i,j=1}^{n}\alpha_{1i}(\alpha_{3j}+\alpha_{3j}^*)K(|x_i|, |x_j|)\Big) \qquad (12)\\
&+\sum_{i,j=1}^{n}(\alpha_{3i}+\alpha_{3i}^*)(\alpha_{3j}+\alpha_{3j}^*)K(|x_i|, |x_j|)\Big) - \frac{1}{2C}\sum_{i=1}^{n}\alpha_{1i}^2\\
&-\frac{1}{2C}\sum_{i=1}^{n}\alpha_{1i}^2 - \frac{1}{2C}\sum_{i=1}^{n}(\alpha_{2i}^2+\alpha_{2i}^{*2})\\
&+\sum_{i=1}^{n}(\alpha_{2i}-\alpha_{2i}^*)y_i + \sum_{i=1}^{n}(\alpha_{3i}-\alpha_{3i}^*)y_i - \sum_{i=1}^{n}(\alpha_{4i}-\alpha_{4i}^*)y_i\\
&+\sum_{i=1}^{n}(\alpha_{3i}+\alpha_{3i}^*)e_i - \sum_{i=1}^{n}(\alpha_{4i}+\alpha_{4i}^*)e_i\Big\}
\end{aligned}$$

subject to

$$\alpha_{1i}, \alpha_{ki}, \alpha_{ki}^* \geq 0, \ k = 2, 3, 4.$$

Solving (12) with the above constraints determines the Lagrange multipliers, $\alpha_{1i}, \alpha_{ki}, \alpha_{ki}^*$. Therefore, the interval nonlinear regression function is given as follows:

$$\begin{aligned}
Y^*(x) = \Big(&\sum_{i=1}^{n}[(\alpha_{2i}-\alpha_{2i}^*)+(\alpha_{3i}-\alpha_{3i}^*)-(\alpha_{4i}-\alpha_{4i}^*)]K(x_i, x),\\
&\sum_{i=1}^{n}[-\alpha_{1i}+2(\alpha_{3i}+\alpha_{3i}^*)-(\alpha_{4i}+\alpha_{4i}^*)]K(|x_i|, |x|),\Big) \qquad (13)\\
Y_*(x) = \Big(&\sum_{i=1}^{n}[(\alpha_{2i}-\alpha_{2i}^*)+(\alpha_{3i}-\alpha_{3i}^*)-(\alpha_{4i}-\alpha_{4i}^*)]K(x_i, x),\\
&\sum_{i=1}^{n}[(\alpha_{3i}+\alpha_{3i}^*)-(\alpha_{4i}+\alpha_{4i}^*)]K(|x_i|, |x|),\Big) \qquad (14)
\end{aligned}$$

References

P. Diamond, H. Tanaka(1998), Fuzzy regression analysis, in: R. Slowinski (Ed.), Fuzzy Sets in Decision Analysis, Operations Research and Statistics, Kluwer, Boston, 349-387.

D.H Hong, C. Hwang(2003), Support vector fuzzy regression machines, Fuzzy Sets and

Systems, 138, 271-281.
H. Tanaka(1987), Fuzzy data analysis by possibilistic linear models, Fuzzy Sets and Systems, 24, 363-375.
H. Tanaka, I. Hayashi, K. Nagasaka(1988), Interval regression analysis by possibilistic measures (in Japanese), The Japanese Journal of Behaviormetrics, 16, 1-7.
H. Tanaka, H. Lee(1998), Interval regression analysis by quadratic programming approach, IEEE Trans. Fuzzy Syst., 6, 473-481.
H. Tanaka, S. Uejima, K. Asia(1992), Linear regression analysis with fuzzy model, IEEE Trans. Man. Cybernet. 12, 903-907.
V. Vapnik(1998), Statistical Learning Theory, Springer, Germany.