Proceedings of the Korean Data And Information Science Society, October 22-23, 2004, pp. 53-59

Estimating Fuzzy Regression with Crisp Input-Output Using Quadratic Loss Support Vector Machine

Changha Hwang[1], Dug Hun Hong[2], Sangbock Lee[3]

Abstract

Support vector machine(SVM) approach to regression can be found in information science literature. SVM implements the regularization technique which has been introduced as a way of controlling the smoothness properties of regression function. In this paper, we propose a new estimation method based on quadratic loss SVM for a linear fuzzy regression model of Tanaka's, and furthermore propose a estimation method for nonlinear fuzzy regression. This approach is a very attractive approach to evaluate nonlinear fuzzy model with crisp input and output data.

Keywords : Crisp data, fuzzy regression, linear programming, quadratic loss, quadratic programming, support vector machine.

1. Introduction

Fuzzy linear regression provides a method for tackling regression problems lacking significance amount of data for determining regression models and with vague relationships between the dependent variables. The concept of fuzzy regression analysis was introduced by Tanaka *et al.*(1982), where an LP based method with symmetric triangular fuzzy parameters was proposed. The method is recommended for practical situations where decisions often have to be made on the basis of imprecise and partially available data where human estimation is influential. This first attempt of applying fuzzy regression was done using non-fuzzy input experimental data. An extension of the idea was reported by Tanaka(1987) comparing the capability to process fuzzy input experimental data. Heshmaty and Kandel(1985) applied this method to forecasting in uncertain

1 Professor, Dept. of Statistical Information, Catholic University of Daegu.
E-mail : chhwang@cu.ac.kr

2 Professor, Dept. of Mathematics, Myongji University

3 Professor, Dept. of Statistical Information, Catholic University of Daegu

environment and Watada(1983) applied the idea of fuzzy regression to fuzzy time-series. Fuzzy data analysis, regarded as a non-statistical procedure for possibilistic systems, was reported by Tanaka(1987), and Tanaka et al.(1982).

Fuzzy regression has been also investigated from the viewpoint of least square regression. Celmins(1987a,b) and Diamond(1988) developed several models for fuzzy least squares fitting. In this paper, we propose a new estimation method based on quadratic loss SVM for a linear fuzzy regression model of Tanaka's, and furthermore propose a estimation method for nonlinear fuzzy regression. Basically, the proposed method utilizes Tanaka and Lee(1998)'s idea based on quadratic programming(QP).

2. Fuzzy Regression for Crisp Input-Output Data

In this section, we illustrate how to get solutions for fuzzy linear regression models using LP and QP approaches proposed by Tanaka et al.(1982) and Tanaka and Lee(1998), respectively.

Suppose that we are given training data $\{(\boldsymbol{x}_i, y_i),\ i=1,\cdots,n\} \subset X \times R$, where X denotes the space of the input patterns. We begin by describing the case of fuzzy linear regression functions $Y(\boldsymbol{x})$, taking the form

$$Y(\boldsymbol{x}) = A_0 + A_1 x_1 + \cdots + A_m x_m = \boldsymbol{A}^t \boldsymbol{x}, \tag{1}$$

where $\boldsymbol{x} = (1, x_1, \cdots, x_m)^t$ is a real input vector, $\boldsymbol{A} = (A_0, A_1, \cdots, A_m)^t$ is an interval coefficient vector, and $Y(\boldsymbol{x})$ is the corresponding estimated interval. An interval coefficient A_i is denoted as $A_i = (a_i, c_i)$ where a_i is a center and c_i is a radius. By interval arithmetic, the regression model (1) can be expressed as

$$\begin{aligned} Y(\boldsymbol{x}_i) &= (a_0, c_0) + (a_1, c_1) x_{i1} + \cdots + (a_m, c_m) x_{im} \\ &= (a_0 + a_1 x_{i1} + \cdots + a_m x_{im}, c_0 + c_1 |x_{i1}| + \cdots + c_m |x_{im}|) \\ &= (\boldsymbol{a}^t \boldsymbol{x}_i, \boldsymbol{c}^t |\boldsymbol{x}_i|), \end{aligned} \tag{2}$$

where $\boldsymbol{a} = (a_0, a_1, \cdots, a_m)^t$, $\boldsymbol{c} = (c_0, c_1, \cdots, c_m)^t$, and $|\boldsymbol{x}_i| = (1, |x_{i1}|, \cdots, |x_{im}|)^t$.

We first illustrate LP approach proposed by Tanaka et al.(1982), which solves the following LP problem

$$J = \sum_{i=1}^{n} \boldsymbol{c}^t |\boldsymbol{x}_i|$$

subject to

$$\boldsymbol{a}^t \boldsymbol{x}_i + (1-h) \boldsymbol{c}^t |\boldsymbol{x}_i| \geq y_i, \quad \boldsymbol{a}^t \boldsymbol{x}_i - (1-h) \boldsymbol{c}^t |\boldsymbol{x}_i| \leq y_i,$$
$$c_i \geq 0, \ i = 0, 1, \cdots, m$$

we illustrate how to get solutions for fuzzy linear regression models using QP approach proposed by Tanaka and Lee(1998). In principle, Tanaka and Lee(1998) use the formulation integrating central tendency and possibilistic property. Thus, we consider a new objective function which reflects both properties of least squares and possibilistic approaches

$$J = \sum_{i=1}^{n}(y_i - \boldsymbol{a}^t\boldsymbol{x}_i)^2 + \sum_{i=1}^{n} \boldsymbol{c}^t|\boldsymbol{x}_i||\boldsymbol{x}_i|^t\boldsymbol{c} \tag{3}$$

where $\sum_{i=1}^{n}|\boldsymbol{x}_i||\boldsymbol{x}_i|^t$ is a symmetric positive definite matrix. Fuzzy linear regression analysis using this new objective function (3) is to determine the interval coefficients $A_i = (a_i, c_i)$, $i = 0, 1, \cdots, m$ by solving the following QP problem:

$$\min_{\boldsymbol{a}, \boldsymbol{c}} \; J = \sum_{i=1}^{n}(y_i - \boldsymbol{a}^t\boldsymbol{x}_i)^2 + \sum_{i=1}^{n} \boldsymbol{c}^t|\boldsymbol{x}_i||\boldsymbol{x}_i|^t\boldsymbol{c} \tag{4}$$

subject to

$$\boldsymbol{a}^t\boldsymbol{x}_i + (1-h)\boldsymbol{c}^t|\boldsymbol{x}_i| \geq y_i, \quad \boldsymbol{a}^t\boldsymbol{x}_i - (1-h)\boldsymbol{c}^t|\boldsymbol{x}_i| \leq y_i,$$
$$c_i \geq 0, \; i = 0, 1, \cdots, m$$

3. Quadratic Loss SVM for Fuzzy 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:

$$\min_{\boldsymbol{a}, \boldsymbol{c}} \; \frac{1}{2}(\|\boldsymbol{a}\|^2 + \|\boldsymbol{c}\|^2) + \frac{\gamma}{2}\left(\sum_{i=1}^{n}\xi_{1i}^2 + \sum_{i=1}^{n}(\xi_{2i}^2 + \xi_{2i}^{*2})\right) \tag{5}$$

subject to

$$\boldsymbol{c}^t|\boldsymbol{x}_i| \leq \xi_{1i}$$
$$y_i - \boldsymbol{a}^t\boldsymbol{x}_i \leq \xi_{2i}, \quad \boldsymbol{a}^t\boldsymbol{x}_i - y_i \leq \xi_{2i}^*, \; i = 1, \cdots, n$$
$$\boldsymbol{a}^t\boldsymbol{x}_i + (1-h)\boldsymbol{c}^t|\boldsymbol{x}_i| \geq y_i, \quad \boldsymbol{a}^t\boldsymbol{x}_i - (1-h)\boldsymbol{c}^t|\boldsymbol{x}_i| \leq y_i,$$
$$c_i \geq 0, \; i = 0, 1, \cdots, m$$

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:

$$
\begin{aligned}
L = & \frac{1}{2}(\| \boldsymbol{a} \|^2 + \| \boldsymbol{c} \|^2) + \frac{\gamma}{2}\left(\sum_{i=1}^{n} \xi_{1i}^2 + \sum_{i=1}^{n}(\xi_{2i}^2 + \xi_{2i}^{*2})\right) \\
& - \sum_{i=1}^{n} \alpha_{1i}(\xi_{1i} - \boldsymbol{d}^t |\boldsymbol{x}_i|) \qquad (6) \\
& - \sum_{i=1}^{n} \alpha_{2i}(\xi_{2i} - y_i + \boldsymbol{a}^t \boldsymbol{x}_i) - \sum_{i=1}^{n} \alpha_{2i}^*(\xi_{2i} - \boldsymbol{a}^t \boldsymbol{x}_i + y_i) \\
& - \sum_{i=1}^{n} \alpha_{3i}(\boldsymbol{a}^t \boldsymbol{x}_i + (1-h)\boldsymbol{c}^t |\boldsymbol{x}_i| - y_i) - \sum_{i=1}^{n} \alpha_{3i}^*(y_i - \boldsymbol{a}^t \boldsymbol{x}_i + (1-h)\boldsymbol{c}^t |\boldsymbol{x}_i|)
\end{aligned}
$$

Here, $\alpha_{1i}, \alpha_{2i}, \alpha_{2i}^*, \alpha_{3i}, \alpha_{3i}^*$ are Lagrange multipliers. It follows from the saddle point condition that the partial derivatives of L with respect to the primal variables $(\boldsymbol{a}, \boldsymbol{c}, \xi_{1i}, \xi_{2i}, \xi_{2i}^*)$ have to vanish for optimality.

$$\frac{\partial L}{\partial \boldsymbol{a}} = 0 \rightarrow \boldsymbol{a} = \sum_{i=1}^{n}(\alpha_{2i} - \alpha_{2i}^*)\boldsymbol{x}_i + \sum_{i=1}^{n}(\alpha_{3i} - \alpha_{3i}^*)\boldsymbol{x}_i \qquad (7)$$

$$\frac{\partial L}{\partial \boldsymbol{c}} = 0 \rightarrow \boldsymbol{c} = -\sum_{i=1}^{n} \alpha_{1i}|\boldsymbol{x}_i| + (1-h)\sum_{i=1}^{n}(\alpha_{3i} + \alpha_{3i}^*)|\boldsymbol{x}_i| \qquad (8)$$

$$\frac{\partial L}{\partial \xi_{1i}} = 0 \rightarrow \xi_{1i} = \frac{1}{\gamma}\alpha_{1i} \qquad (9)$$

$$\frac{\partial L}{\partial \xi_{2i}^{(*)}} = 0 \rightarrow \xi_{2i}^{(*)} = \frac{1}{\gamma}\alpha_{2i}^{(*)} \qquad (10)$$

Substituting (7)-(10) into (6) yields the dual optimization problem.

$$
\begin{aligned}
\text{maximize}\Bigg\{ & -\frac{1}{2}\Bigg(\sum_{i,j=1}^{n}(\alpha_{2i} - \alpha_{2i}^*)(\alpha_{2j} - \alpha_{2j}^*)\boldsymbol{x}_i^t \boldsymbol{x}_j \\
& + 2\sum_{i,j=1}^{n}(\alpha_{2i} - \alpha_{2i}^*)(\alpha_{3j} - \alpha_{3j}^*)\boldsymbol{x}_i^t \boldsymbol{x}_j + \sum_{i,j=1}^{n}(\alpha_{3i} - \alpha_{3i}^*)(\alpha_{3j} - \alpha_{3j}^*)\boldsymbol{x}_i^t \boldsymbol{x}_j \\
& + \sum_{i,j=1}^{n} \alpha_{1i}\alpha_{1j}|\boldsymbol{x}_i|^t|\boldsymbol{x}_j| - 2(1-h)\sum_{i,j=1}^{n} \alpha_{1i}(\alpha_{3j} + \alpha_{3j}^*)|\boldsymbol{x}_i|^t|\boldsymbol{x}_j| \qquad (11) \\
& + (1-h)^2\sum_{i,j=1}^{n}(\alpha_{3i} + \alpha_{3i}^*)(\alpha_{3j} + \alpha_{3j}^*)|\boldsymbol{x}_i|^t|\boldsymbol{x}_j|\Bigg) - \frac{1}{2\gamma}\sum_{i=1}^{n}\alpha_{1i}^2 \\
& - \frac{1}{2\gamma}\sum_{i=1}^{n}(\alpha_{2i}^2 + \alpha_{2i}^{*2}) + \sum_{i=1}^{n}(\alpha_{2i}^2 - \alpha_{2i}^{*2})y_i + \sum_{i=1}^{n}(\alpha_{3i}^2 - \alpha_{3i}^{*2})y_i\Bigg\}
\end{aligned}
$$

subject to

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

Solving (11) with above constraints determines the Lagrange multipliers, α_{1i}, $\alpha_{ki}, \alpha_{ki}^*$. Hence, if $\boldsymbol{c}^t|\boldsymbol{x}| \geq 0$, then the linear interval regression function is as follows:

$$Y(\boldsymbol{x}) = (\boldsymbol{a}^t \boldsymbol{x}, \boldsymbol{c}^t|\boldsymbol{x}|) \qquad (12)$$

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 x_i by a map $\Phi: R^d \to E$ 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 (11) is in the form of inner products $x_i^t x_j$, $|x_i|^t |x_j|$. 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

$$\begin{aligned}
\text{maximize}\Big\{ & -\frac{1}{2}\Big(\sum_{i,j=1}^{n} (\alpha_{2i}-\alpha_{2i}^{*})(\alpha_{2j}-\alpha_{2j}^{*})K(x_i, x_j) \\
& + 2\sum_{i,j=1}^{n} (\alpha_{2i}-\alpha_{2i}^{*})(\alpha_{3j}-\alpha_{3j}^{*})K(x_i, x_j) \\
& + \sum_{i,j=1}^{n} (\alpha_{3i}-\alpha_{3i}^{*})(\alpha_{3j}-\alpha_{3j}^{*})K(x_i, x_j) + \sum_{i,j=1}^{n} \alpha_{1i}\alpha_{1j}K(|x_i|, |x_j|) \\
& - 2(1-h)\sum_{i,j=1}^{n} \alpha_{1i}(\alpha_{3j}+\alpha_{3j}^{*})K(|x_i|, |x_j|) \qquad (13) \\
& + (1-h)^2 \sum_{i,j=1}^{n} (\alpha_{3i}+\alpha_{3i}^{*})(\alpha_{3j}+\alpha_{3j}^{*})K(|x_i|, |x_j|)\Big) - \frac{1}{2\gamma}\sum_{i=1}^{n} \alpha_{1i}^2 \\
& - \frac{1}{2\gamma}\sum_{i=1}^{n} (\alpha_{2i}^2+\alpha_{2i}^{*2}) + \sum_{i=1}^{n} (\alpha_{2i}^2-\alpha_{2i}^{*2})y_i + \sum_{i=1}^{n} (\alpha_{3i}^2-\alpha_{3i}^{*2})y_i \Big\}
\end{aligned}$$

subject to

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

4. Empirical Study

In this section, one example is used to verify the effectiveness of the proposed SVM for the fuzzy regression with crisp data. This simulation was conducted in the Matlab environment. In this study, we show the results for fuzzy nonlinear regression only. The data sets are given in Table 1 below, which are taken from Gunn(1998). Gunn(1998) shows that the nonlinear model is appropriate for this data set. Here, we use Gaussian kernel and penalty constant $\gamma = 35$, kernel parameters $\sigma = 0.8$, which have been determined by the leave-one-out method.

Table 1: The data set for fuzzy nonlinear regression

x_i	1	3	4	5.6	7.8	10.2	11.0	11.5	12.7
y_i	-1.6	-1.8	-1.0	1.2	2.2	6.8	10.0	10.0	10.0

In Figure 1 the solid, the dash-dotted and the dotted lines represent the fitted center, the upper and the lower limitations, respectively, when $h = 0$. As seen from Figure 1, the proposed method works quite well. In fact, this method is much simpler and computationally inexpensive. Our procedure has an advantage that we do not need assume the underlying model function.

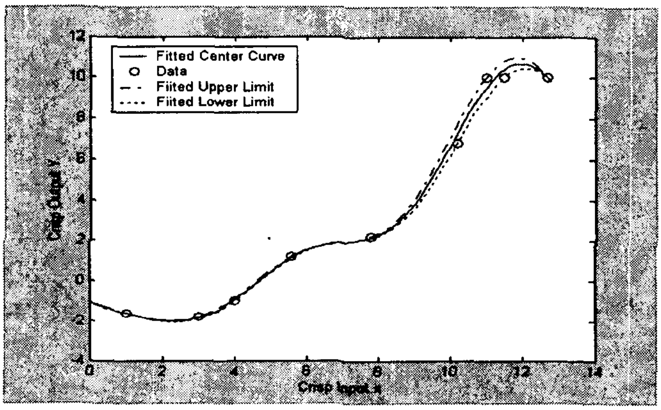

Figure 1. The estimation result for fuzzy nonlinear regression

References

1. Celmins, A. (1987a). Least model fitting to fuzzy vector data, *Fuzzy Sets and Systems*, 22, 245-269.
2. Celmins, A. (1987b). Multidimensional least-squares fitting of fuzzy models, *Math. Modelling*, 9, 669-690.
3. Diamond, P. (1988). Fuzzy least squares, *Inform. Sci.*, 46, 141-157.
4. Gunn, S. (1998). Support vector machines for classification and regression, ISIS Technical Report, U. of Southampton.
5. Heshmaty, B., Kandel, A. (1985). Fuzzy linear regression and its applications to forecasting in uncertain environment, *Fuzzy Sets and Systems*, 15, 159-191.
6. Tanaka, H., Uejima, S., Asia, K. (1982). Linear regression analysis with fuzzy model, *IEEE Trans. Man. Cybernet*, 12, 903-907.

7. Tanaka, H. (1987). Fuzzy data analysis by possibilistic linear models, *Fuzzy Sets and Systems,* 24, 363-375.
8. Tanaka, H., Lee, H. (1998). Interval regression analysis by quadratic programming approach, IEEE Trans. Fuzzy Systems, 6, 473-481.
9. Vapnik, V. (1995). The Nature of Statistical Learning Theory, Springer, New York.
10. Watada, J. (1983). Theory of fuzzy multivariate analysis and its applications, Ph. D. Dissertation, University of Osaka Prefecture.