Engineering Applications of Artificial Intelligence 47 (2016) 16–24

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Optimizing h value for fuzzy linear regression with asymmetric triangular fuzzy coefficients

Fangning Chen, Yizeng Chen, Jian Zhou*, Yuanyuan Liu

School of Management, Shanghai University, Shanghai 200444, China

ARTICLE INFO

Available online 18 March 2015

Keywords:
Fuzzy linear regression
h value
System credibility
Asymmetric triangular fuzzy number
Crisp input–output

ABSTRACT

The parameter h in a fuzzy linear regression model is vital since it influences the degree of the fitting of the estimated fuzzy linear relationship to the given data directly. However, it is usually subjectively pre-selected by a decision-maker as an input to the model in practice. In Liu and Chen (2013), a new concept of system credibility was introduced by combining the system fuzziness with the system membership degree, and a systematic approach was proposed to optimize the h value for fuzzy linear regression analysis using the minimum fuzziness criterion with symmetric triangular fuzzy coefficients. As an extension, in this paper, their approach is extended to asymmetric cases, and the procedure to find the optimal h value to maximize the system credibility of the fuzzy linear regression model with asymmetric triangular fuzzy coefficients is described. Some illustrative examples are given to show the detailed procedure of this approach, and comparative studies are also conducted via the testing data sets.

1. Introduction

Fuzzy regression analysis was proposed by Tanaka et al. (1982) for the linear case using the fuzzy functions defined by Zadeh's extension principle (Zadeh, 1975), in which the observed values can differ from the estimated values to a certain degree of belief. This method is recommended for practical situations where decisions often have to be made on the basis of imprecise and/or partially available data due to human estimations. Fuzzy regression analysis is a powerful tool in many decision domains in estimating relationships among variables with fuzzy, incomplete information (Chen et al., 2004; Höglund, 2013).

Generally, according to criterion functions, the existing fuzzy linear regression (FLR) methods can be roughly classified into two categories: FLR methods using the fuzzy least-squares criterion and FLR methods using the minimum fuzziness criterion. The first category of FLR methods aims to minimize the sum of squared errors in the estimated value based on the notion of distance between the predicted fuzzy outputs and the observed fuzzy outputs, which is indeed a fuzzy extension of the ordinary least squares. The second category of FLR methods, namely the minimum fuzziness approach, aims to build fuzzy linear models by minimizing the system fuzziness subject to including the data points of each sample within a specified feasible data interval. In this field, most of results focus on improving the linear regression method with the initial setting of symmetric triangular coefficients and the assumption of crisp input–output data in Tanaka et al. (1982). For example, many FLR models with some other types of fuzzy coefficients were proposed in the literature including asymmetric triangular fuzzy coefficients, symmetric fuzzy coefficients, LR-type coefficients, trapezoidal fuzzy coefficients, and exponential fuzzy coefficients (see, e.g., Ge and Wang, 2007; Kheirfam and Verdegay, 2013; Tanaka, 1997; Tanaka et al., 1995; Yen et al., 1999).

In the FLR methods using the minimum fuzziness criterion, there is an important input parameter h, which refers to the degree of the fitting of the estimated fuzzy linear model to the given data, and is subjectively pre-selected by a decision-maker in real applications. The selection of a suitable h value for the FLR model is very vital since it determines the range of the possibility distributions of the obtained fuzzy coefficients directly. Actually, a higher h value will produce a large but unnecessary spread values of fuzzy coefficients, which has no operational definition or interpretation, while a lower h value may lead to a very narrow predictive interval so that the reliability of the FLR model is doubtable (Liu and Chen, 2013). In 1988, Tanaka and Watada (1988) first discussed the selection of a proper h value for the FLR model. They advised that a greater h value should be introduced if the observed data pairs are relative small. Besides, Bárdossy (1990) stated that the selection of the h value could be generally based upon the decision-maker's belief in the FLR model, and then recommended an h value between 0.5 and 0.7. Since these criteria are ad hoc, vague, and rather difficult to be justified or applied in

* Corresponding author. Tel.: +86 21 66134414 805.
E-mail address: zhou_jian@shu.edu.cn (J. Zhou).

http://dx.doi.org/10.1016/j.engappai.2015.02.011

real situations, Moskowitz and Kim (1993) studied and examined the relationship among the h value, the membership function shape, and the spreads of fuzzy parameters in FLR with symmetric fuzzy numbers. Subsequently they developed a systematic approach to assessing a proper h value for the FLR model, which should satisfy the decision-maker's beliefs regarding the shape and range of the possibility distributions of fuzzy coefficients. Considering the situations that the observed data set contains a considerable level of noise or uncertainty, Ge and Wang (2007) suggested that the h value for the FLR model should be inversely proportional to the input noise. Besides, Shakouri and Nadimi (2009) introduced a measuring index of inequality of fuzzy numbers and proposed a programming with a new objective function as well as constraints to obtain an optimal h based on an idea of reducing the distance between the output of the possibilistic model and the measured output. Recently, Liu and Chen (2013) formulated a novel approach to optimizing the h value for FLR using the minimum fuzziness criterion based on a new notion of credibility, which may evaluate the reliability of FLR models. They focused on improving the classical and widely accepted FLR model proposed by Tanaka et al. (1982), and proposed a simple and easy-understood calculation process for the optimal h value by taking into account both the system fuzziness and the system membership degree. In their study, the given data are crisp input–output, and the coefficients are assumed to be symmetric triangular fuzzy numbers (TFNs). Considering that there are a great deal of data sets that generate scatter plots in which the data do not fall symmetrically on both sides of the regression line in practice, in the present paper, the approach in Liu and Chen (2013) is extended to asymmetric cases. The coefficients in the FLR model are assumed as asymmetric TFNs, and then an approach is presented for finding the optimal h value to maximize the reliability of the FLR model. Some comparative studies are also conducted.

The rest of the paper is organized as follows. In Section 2, Tanaka's FLR method with asymmetric TFNs is described. In Section 3, the concept of credibility is introduced to measure the performance of FLR models with different h values. In Section 4, a systematic approach is formulated to optimize the h value for FLR models with asymmetric TFNs. In Section 5, some numerical simulations are used to demonstrate the performance of the proposed approach.

2. FLR with asymmetric triangular fuzzy coefficients

In FLR analysis, the explained variable is assumed to be a linear combination of the explanatory variables. This relationship should be obtained from a sample of m observations $\{(y_1,\boldsymbol{x}_1),(y_2,\boldsymbol{x}_2),\dots,(y_i,\boldsymbol{x}_i),\dots,(y_m,\boldsymbol{x}_m)\}$, where y_i is the ith observed crisp output, and $\boldsymbol{x}_i=(x_{i0},x_{i1},\dots,x_{ij},\dots,x_{in})$ is the ith observed crisp input vector. Moreover, $x_{i0}=1$ for all i, and x_{ij} is the observed value for the jth variable in the ith case of the sample. In particular, the fuzzy linear function has to be estimated as follows:

$$\tilde{y}_i=f(\boldsymbol{x}_i)=\tilde{A}_0+\tilde{A}_1x_{i1}+\cdots+\tilde{A}_jx_{ij}+\cdots+\tilde{A}_nx_{in}=\sum_{j=0}^{n}\tilde{A}_jx_{ij} \tag{1}$$

where $\tilde{y}_i$ is the fuzzy estimation of y_i, and $\tilde{A}_j$, $j=0,1,\dots,n$, are fuzzy coefficients in terms of symmetric or asymmetric fuzzy numbers, which can be determined by solving an FLR model with symmetric fuzzy numbers or asymmetric fuzzy numbers.

When the FLR with symmetric coefficients is applied, the obtained regression line may not be the best fit because of the existence of a large number of outliers and high residuals. In practice, there are a great deal of data sets that generate scatter plots in which the data do not fall symmetrically on both sides of the regression line. Thus, the FLR with symmetric TFNs was extended asymmetrically by Yen et al. (1999) as follows.

If $\tilde{A}_j$ in (1) has an asymmetric triangular membership function, it can be uniquely defined by a triplet $\tilde{A}_j=(a_j^L,a_j^C,a_j^R)$, where a_j^C is the centre value, and a_j^L and a_j^R are the left and right spreads of $\tilde{A}_j$, respectively. The goal in FLR is to determine $f(\boldsymbol{x}_i)$ by minimizing the system fuzziness subject to the following inclusion conditions (Tanaka et al., 1982):

$$y_i\in[f(\boldsymbol{x}_i)]^h,\quad i=1,2,\dots,m. \tag{2}$$

Here h $(0\le h<1)$ is a parameter predetermined subjectively by the design team according to their engineering knowledge, and $[f(\boldsymbol{x}_i)]^h$ is the h-level set of the predicted fuzzy output $\tilde{y}_i=f(\boldsymbol{x}_i)$ from the FLR model in (1) corresponding to the input vector $\boldsymbol{x}_i$, which is an interval. Since the fuzzy coefficients $\tilde{A}_j, j=0,1,\dots,n$, in (1) are all asymmetric TFNs, according to fuzzy arithmetic on TFNs, the predicted fuzzy output $f(\boldsymbol{x}_i)$ from the FLR model in (1) is also calculated as an asymmetric TFN. In order to further provide the expression of $[f(\boldsymbol{x}_i)]^h$, we define

$$x_{ij}^{+}=\begin{cases}x_{ij} & \text{if } x_{ij}\ge 0\\ 0 & \text{otherwise}\end{cases} \tag{3}$$

and

$$x_{ij}^{-}=\begin{cases}0 & \text{if } x_{ij}\ge 0\\ -x_{ij} & \text{otherwise}\end{cases} \tag{4}$$

for $i=1,2,\dots,m$ and $j=0,1,\dots,n$, respectively. It is obvious that x_{ij}^{+} and x_{ij}^{-} are both nonnegative and satisfy that

$$x_{ij}=x_{ij}^{+}-x_{ij}^{-}\quad\text{and}\quad |x_{ij}|=x_{ij}^{+}+x_{ij}^{-}. \tag{5}$$

Thus, if we denote the predicted fuzzy output $f(\boldsymbol{x}_i)$ via its left spread $f^L(\boldsymbol{x}_i)$, peak point $f^C(\boldsymbol{x}_i)$ and right spread $f^R(\boldsymbol{x}_i)$ as $f(\boldsymbol{x}_i)=(f^L(\boldsymbol{x}_i),f^C(\boldsymbol{x}_i),f^R(\boldsymbol{x}_i))$, then by the sum and product operations of TFNs, we can obtain $f^L(\boldsymbol{x}_i)$, $f^C(\boldsymbol{x}_i)$ and $f^R(\boldsymbol{x}_i)$, respectively, as

$$f^L(\boldsymbol{x}_i)=\sum_{j=0}^{n}a_j^Lx_{ij}^{+}+\sum_{j=0}^{n}a_j^Rx_{ij}^{-}, \tag{6}$$

$$f^C(\boldsymbol{x}_i)=\sum_{j=0}^{n}a_j^Cx_{ij}, \tag{7}$$

$$f^R(\boldsymbol{x}_i)=\sum_{j=0}^{n}a_j^Rx_{ij}^{+}+\sum_{j=0}^{n}a_j^Lx_{ij}^{-}. \tag{8}$$

Besides, the h-level $(0\le h<1)$ set of $f(\boldsymbol{x}_i)=(f^L(\boldsymbol{x}_i),f^C(\boldsymbol{x}_i),f^R(\boldsymbol{x}_i))$ is calculated as the following interval:

$$[f(\boldsymbol{x}_i)]^h=[f^C(\boldsymbol{x}_i)-(1-h)f^L(\boldsymbol{x}_i),f^C(\boldsymbol{x}_i)+(1-h)f^R(\boldsymbol{x}_i)]. \tag{9}$$

As a result, the inclusion relation in (2) can be rewritten as

$$f^C(\boldsymbol{x}_i)-(1-h)f^L(\boldsymbol{x}_i)\le y_i\le f^C(\boldsymbol{x}_i)+(1-h)f^R(\boldsymbol{x}_i),\quad i=1,2,\dots,m. \tag{10}$$

Furthermore, the system fuzziness, denoted by Δ, is defined by Tanaka et al. (1982) as the total covering area of predicted fuzzy outputs, i.e.

$$\Delta=\sum_{i=1}^{m}\Delta\tilde{y}_i=\sum_{i=1}^{m}\frac{1}{2}(f^L(\boldsymbol{x}_i)+f^R(\boldsymbol{x}_i))=\frac{1}{2}\sum_{i=1}^{m}\sum_{j=0}^{n}(a_j^L+a_j^R)(x_{ij}^{+}+x_{ij}^{-}) \tag{11}$$

in which $\Delta\tilde{y}_i$ is the fuzziness with respect to $\tilde{y}_i$ in the asymmetric case, and can be given as

$$\Delta\tilde{y}_i=\frac{1}{2}\sum_{j=0}^{n}(a_j^L+a_j^R)(x_{ij}^{+}+x_{ij}^{-}). \tag{12}$$