Applied Soft Computing xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Applied Soft Computing

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

A hybrid fuzzy regression model and its application in hydrology engineering

Q1 J. Chachi [a,*], S.M. Taheri [b], N.R. Arghami [c]

[a] Department of Mathematics, Statistics and Computer Sciences, Semnan University, Semnan 35195-363, Iran
[b] Department of Engineering Science, College of Engineering, University of Tehran, P.O. Box 11365-4563, Tehran, Iran
[c] Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Iran

ARTICLE INFO

Article history:
Received 8 January 2013
Received in revised form 23 May 2014
Accepted 12 September 2014
Available online xxx

Keywords:
Discharge
Fuzzy regression
Hydrology
Multivariate adaptive regression splines (MARS)
Spline basis function
Suspended load

ABSTRACT

The goal of this paper is to handle the large variation issues in fuzzy data by constructing a variable spread multivariate adaptive regression splines (MARS) fuzzy regression model with crisp parameters estimation and fuzzy error terms. It deals with imprecise measurement of response variable and crisp measurement of explanatory variables. The proposed method is a two-phase procedure which applies the MARS technique at phase one and an optimization problem at phase two to estimate the center and fuzziness of the response variable. The proposed method, therefore, handles two problems simultaneously: the problem of large variation issue and the problem of variation spreads in fuzzy observations. A realistic application of the proposed method is also presented, by which the suspended load is modeled using discharge in a hydrology engineering problem. Empirical results demonstrate that the proposed approach is more efficient and more realistic than some well-known least-squares fuzzy regression models.

1. Introduction

Fuzzy regression analysis is a widely known method for seeking the fuzzy relationship between input variables (also known as independent or explanatory variables) and output variable (also called dependent or response variable) based on a crisp (exact) or fuzzy (imprecise) data set. Two main approaches to construct a regression model in fuzzy environments are:

1. the possibilistic approach, e.g. see [2,3,6,32,38,44], and
2. the fuzzy least-squares approach, e.g. see [4,8,10,11,13,14,33,34,41,42].

However, during recent years, it is shown that in some cases the combined techniques, which integrate several single methods, have greater accuracy than any individual method [19,27,30]. Although it is shown that these methods have some theoretical and applied advantages the issue of the large variation data however has not been discussed in these studies. The problem of large variation data, specially, occurs in the analysis of some real-world data, such as: incomes data, rainfall data, data related to the characteristics of soil, and so on.

To resolve the above problem, nonparametric models can be used. One of the most promising nonparametric techniques is multivariate adaptive regression splines (MARS) which models relationships that are nearly additive or involve interactions with fewer variables [16]. It essentially builds flexible models by fitting piecewise linear regressions; that is, the non-linearity of a model is approximated through the use of separate regression slopes in different intervals of the variable space [18]. Over the last years, MARS has been compared with a number of parametric and nonparametric approximation routines in terms of its accuracy, efficiency, robustness, model transparency, and simplicity [1,24–26,37].

In this paper, MARS technique and a mathematical programming method are integrated to propose a new hybrid fuzzy regression procedure to cope with the problem of modeling and analyzing the large variation data. We consider the case when the response variable is fuzzy and the explanatory variables are crisp. This situation commonly arises in practical studies especially in hydrology engineering (as we shall see in Section 6).

The proposed method is a two-phase procedure for computation of fuzzy regression that is simple and gives good solutions. At the first phase, this correspondence employs the MARS technique to estimate the crisp regression coefficients of the model using the defuzzified values of fuzzy observations of response and the

* Corresponding author. Tel.: +98 231 3366292; fax: +98 231 3354082.
E-mail addresses: jchachi@profs.semnan.ac.ir, j.chachi@math.iut.ac.ir (J. Chachi).

http://dx.doi.org/10.1016/j.asoc.2014.09.010
1568-4946/© 2014 Published by Elsevier B.V.

crisp observations of explanatory variables. At the second phase, using the evaluation criterion proposed by Hojati et al. [19] as the objective function of a programming problem, a mathematical programming model is then constructed to determine the fuzzy error term associated with each observation. The model is a variable spread model, and so can avoid the spread increasing problem. A real-life problem in hydrology engineering is used to illustrate the applicability of the introduced method. The performance of the proposed approach with respect to some well-known fuzzy regression models are considered in a comparative study.

The structure of the paper is as follows. In the section below, we shall describe some aspects of fuzzy set, fuzzy arithmetic and multivariate adaptive regression splines known as MARS. In Section 3, using MARS technique, we describe a variable spreads MARS-fuzzy regression model for fuzzy response and crisp explanatory variables. In Section 4, forecasting via fuzzy inference system is illustrated. Two goodness of fit criteria are recalled in Section 5 for evaluating the fuzzy regression models. In Section 6, we explain the applicability of the proposed hybrid model to estimate the suspended load based on the discharge, when the available real data of the suspended load are imprecise (fuzzy) rather than crisp (exact). Finally, we state important conclusions of the paper in Section 7.

2. Preliminaries

2.1. Fuzzy sets and fuzzy arithmetic

A fuzzy set $\tilde{A}$ on the universal set $\mathbb{X}$ is described by its membership function $\tilde{A}(x) : \mathbb{X} \to [0, 1]$. In this paper, we assume that $\mathbb{X} = \mathbb{R}$, the set of real numbers. The crisp set $A_\alpha = \{x \in \mathbb{R} : \tilde{A}(x) \geq \alpha\}$, $\alpha \in (0, 1]$, is called the α-cut of $\tilde{A}$, and for $\alpha = 0$ we assume $A_0 = \mathrm{cl}\{x \in \mathbb{R} : \tilde{A}(x) > 0\}$, where *cl* is the closure operator.

A specific class of fuzzy sets on $\mathbb{R}$ is the so-called *LR*-fuzzy numbers $\tilde{N} = (n, l, r)_{LR}$ with central value $n \in \mathbb{R}$, left and right spreads $l \in \mathbb{R}^+$, $r \in \mathbb{R}^+$, decreasing left and right shape functions $L : \mathbb{R}^+ \to [0, 1]$, $R : \mathbb{R}^+ \to [0, 1]$, with $L(0) = R(0) = 1$. Typically, the *LR*-fuzzy number $\tilde{N}$ has the following membership function [45]

$$\tilde{N}(x) = \begin{cases} L\left(\dfrac{n-x}{l}\right) & \text{if} \quad x \leq n, \\ R\left(\dfrac{x-n}{r}\right) & \text{if} \quad x > n. \end{cases} \tag{1}$$

A special type of *LR*-fuzzy number is the so-called triangular fuzzy number, denoted by $\tilde{N} = (n, l, r)_T$. The membership function of triangular fuzzy number $\tilde{N}$ is as follow

$$\tilde{N}(x) = \begin{cases} \dfrac{x-(n-l)}{l} & \text{if} \quad x \in [n-l, n], \\ \dfrac{(n+r)-x}{r} & \text{if} \quad x \in (n, n+r], \\ 0 & \text{if} \quad x \notin [n-l, n+r]. \end{cases} \tag{2}$$

For $l = r$, the triangular fuzzy number $\tilde{N}$ is called symmetric triangular fuzzy number and is abbreviated by $\tilde{N} = (n, l)_T$.

For the algebraic operations of *LR*-fuzzy numbers, we have the following result on the basis of Zadeh's extension principle (for more details, see [45]). Let $\tilde{M} = (m, l_m, r_m)_{LR}$ and $\tilde{N} = (n, l_n, r_n)_{LR}$ be two *LR*-fuzzy numbers and λ be a real number. Then

$$\lambda \otimes \tilde{M} = \begin{cases} (\lambda m, \lambda l_m, \lambda r_m)_{LR} & \text{if} \quad \lambda > 0, \\ \mathcal{I}_{\{0\}} & \text{if} \quad \lambda = 0, \\ (\lambda m, |\lambda| r_m, |\lambda| l_m)_{RL} & \text{if} \quad \lambda < 0, \end{cases} \tag{3}$$

$$\lambda \oplus \tilde{M} = (\lambda + m, l_m, r_m)_{LR}, \tag{4}$$

$$\tilde{M} \oplus \tilde{N} = (m + n, l_m + l_n, r_m + r_n)_{LR}, \tag{5}$$

where $\mathcal{I}_A$ stands the characteristic function of a crisp set A.

Defuzzification: Many defuzzification approaches have been proposed in the literature, of which the center of gravity (COG) method (also called the centroid method) is the most common used method [5,35]. The COG method calculates N_c, the defuzzified value of fuzzy number $\tilde{N}$, as follows:

$$N_c = \frac{\int x\tilde{N}(x)\,dx}{\int \tilde{N}(x)\,dx}. \tag{6}$$

We can easily obtain the defuzzified value of $\tilde{N} = (n, l, r)_T$ as $N_c = 1/3(3n - l + r)$.

2.2. Multivariate adaptive regression splines (MARS): a brief review

Multivariate adaptive regression splines (MARS) is a non-parametric regression modeling procedure which was first introduced by Friedman [16] to efficiently approximate the relationship between a dependent variable (y) and a set of independent variables ($\mathbf{x}$) in a piecewise regression, especially when the data set is large and/or the relationships between the variables does not follow a linear function. The MARS model takes the form of an expansion in multivariate spline basis functions

$$y = \beta_0 + \sum_{m=1}^{M} \beta_m B_m(\mathbf{x}), \tag{7}$$

where $\beta_0, \beta_1, \ldots, \beta_M$, are the coefficients of the basis functions determined by a least-squares regression, M is the number of basis functions, and

$$B_m(x_1, \ldots, x_n) = \prod_{i=1}^{K_m} \max\{S_{im}(x_{j(i,m)} - t_{im}), 0\}, \tag{8}$$

where $S_{im} = \pm 1$, $x_{j(i,m)}$'s are the explanatory variables associated with the basis function $B_m(x_1, \ldots, x_n)$, i.e. the values of jth explanatory variables at ith node of mth basis function, K_m is the level of interaction between $j(i, m)$ variables, and t_{im} indicates the knot locations for $B_m(x_1, \ldots, x_n)$ [16,18].

Basis functions are selected from the collection $\mathcal{C}$ where

$$\mathcal{C} = \{\{\max\{(x_j - t), 0\}, \max\{(t - x_j), 0\}\}_{t,j} : t \in \{x_{j1}, \ldots, x_{jn}\}, j = 1, \ldots, k\}. \tag{9}$$

Each function is piecewise linear with a knot t at every x_{ji}. One might assume that only piecewise linear functions can be formed from basis functions, but basis functions can be multiplied together to form non-linear functions, which are either order one or cubic, depending on the degree of continuity of the approximation. Therefore, by allowing the basis function to bend at the knots, MARS can model functions that differ in behavior over the domain of each variable [18,37].

MARS models are developed through a two-stage forward/backward stepwise regression procedure which finds the location and number of the needed spline basis functions. To make the MARS algorithm computationally affordable, the maximum number of basis functions, the maximum number of knots considered, the minimum number of observations between knots, and the highest order of interaction terms in the model are specified by the analyst (for more details see [18]). Generally, finding a MARS model is computationally complex. In this paper, we use "`earth`" package in the software "`R-2.14.1`" to report the numerical results [15].

Please cite this article in press as: J. Chachi, et al., A hybrid fuzzy regression model and its application in hydrology engineering, Appl. Soft Comput. J. (2014), http://dx.doi.org/10.1016/j.asoc.2014.09.010