Engineering Applications of Artificial Intelligence 64 (2017) 286–294

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Probabilistic fuzzy regression approach for preference modeling

Huimin Jiang, C.K. Kwong *, Woo-Yong Park [1]

Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong, China

ARTICLE INFO

Keywords:
Probabilistic fuzzy regression
Preference modeling
Chaos optimization algorithm

ABSTRACT

Two types of uncertainty, namely, randomness and fuzziness, exist in preference modeling. Fuzziness is mainly caused by human subjective judgment and incomplete knowledge, and randomness often originates from the variability of influences on the inputs and outputs of a preference model. Various techniques have been utilized to develop preference models. However, only few previous studies have addressed both fuzziness and randomness in preference modeling. Among these limited studies, none have considered the randomness caused by particular independent variables. To fill this research gap, this study proposes probabilistic fuzzy regression (PFR), a new approach for preference modeling. PFR considers both the fuzziness of data sets and the randomness caused by independent variables. In the proposed approach, probability density functions (PDFs) are adopted to model randomness. The parameter settings of the PDFs are determined using a chaos optimization algorithm. The probabilistic terms of the PFR models are generated according to the expected value functions of the random variables. Fuzzy regression analysis is employed to determine the fuzzy coefficients for all the terms of the PFR models. An industrial case study of a tea maker design is used to illustrate the applicability of PFR and evaluate its effectiveness. Modeling results obtained from PFR are compared with those obtained from statistical regression, fuzzy regression, and fuzzy least-squares regression. Results of the training and validation tests show that PFR outperforms the other approaches in terms of training and validation errors.

1. Introduction

Survey/experimental data is often used to develop empirical models that relate the inputs and outputs of a system or process. Various approaches for developing empirical models have been attempted. These approaches include quantification theory I (Chang, 2008), ordinal logistic regression (Barone et al., 2007), artificial neural networks (Lai et al., 2005), fuzzy logic approach (Lau et al., 2006), multiple statistical regression (Han et al., 2000), fuzzy linear regression (Sekkeli et al., 2010), particle swarm optimization-based fuzzy regression (Chan et al., 2011a), neural fuzzy systems (Kwong et al., 2009), kernel-based nonlinear fuzzy regression (Su et al., 2013), fuzzy polynomial regression based on fuzzy neural networks (Otadi, 2014), fuzzy regression models using fuzzy distances (de Hierro et al., 2016), and fuzzy regression models based on least absolute deviation (Li et al., 2016). Development of empirical models using survey/experimental data often involves both fuzziness and randomness. Fuzziness is mainly caused by human subjective judgment and incomplete knowledge, and randomness often originates from the variability of influences on the inputs and outputs of a system or process. Only few previous studies have examined both fuzziness and randomness in empirical modeling. Watada et al. (2009) proposed a confidence-interval-based fuzzy random regression approach to address the uncertainties caused by fuzziness and randomness in modeling. In their study, variables were regarded as known fuzzy numbers and probabilities. Kwong et al. (2008) proposed a fuzzy least-squares regression approach to capture fuzziness and randomness simultaneously in modeling manufacturing processes. However, the approach does not specifically address the randomness caused by independent variables.

Preference modeling is aimed at developing models to relate customer preferences and design parameters where customer surveys are commonly adopted to understand customers' preferences and the survey results are used to generate preference models. A number of studies have been conducted to develop preference models via survey and experimental data. Various statistical techniques, such as partial least squares analysis (Nagamachi, 2008) and statistical linear regression (Han et al., 2000; You et al., 2006) have been adopted to model customer preference. However, in customer surveys, customers' responses are always imprecise such as "quite good" and "not very well". Thus, survey results unavoidably contain a high degree of fuzziness. Numerous fuzzy

* Corresponding author.
E-mail address: c.k.kwong@polyu.edu.hk (C.K. Kwong).
[1] His name changed from Jin-Kyu Park to Woo-Yong Park in 2013.

http://dx.doi.org/10.1016/j.engappai.2017.06.019
Received 23 May 2016; Received in revised form 22 May 2017; Accepted 22 June 2017
0952-1976/

approaches for preference modeling have been employed to address the fuzziness in preference modeling. These approaches include fuzzy inference techniques (Liu et al., 2007; Fung et al., 1999), fuzzy rule-based approach (Lau et al., 2006; Park and Han, 2004; Fung et al., 1998) fuzzy logic approach (Lin et al., 2007), fuzzy linear regression (Sekkeli et al., 2010; Shimizu and Jindo, 1995; Chen et al., 2004), nonlinear programming-based fuzzy regression (Chen and Chen, 2006), genetics-based fuzzy regression (Chan et al., 2011b), chaos-based fuzzy regression (Jiang et al., 2013), a stepwise-based fuzzy regression (Chan et al., 2015), and a forward selection-based fuzzy regression (Chan and Ling, 2016). However, all these techniques can only be utilized to deal with either randomness or fuzziness in preference modeling. Kwong et al. (2010) proposed a generalized fuzzy least-squares regression approach to address both fuzziness and randomness in preference modeling. In their proposed approach, Kwong et al. assumed that the estimation error is random and the objective function minimizes the sum of the squares of the residual error (Chang, 2001). However, the approach does not consider the randomness caused by independent variables.

In the current study, a new approach to preference modeling, namely, probabilistic fuzzy regression (PFR), is proposed. PFR can address the fuzziness caused by human subjective judgment and the randomness caused by random variables. Probability density functions (PDFs) are adopted in the proposed approach to model the randomness of independent (random) variables. A chaos optimization algorithm (COA) is employed to determine the parameter settings of the PDFs, and PDFs are then generated. The expected value functions of the random variables based on the PDFs are then generated and incorporated into the PFR models. Fuzzy regression analysis is then conducted to determine the fuzzy coefficients for all the terms of the PFR model.

The remainder of the paper is organized as follows. Section 2 presents the proposed PFR. Section 3 describes a case study on modeling consumer preference based on the proposed approach. Section 4 presents the validation of the proposed approach, and Section 5 provides the conclusions.

2. Probabilistic fuzzy regression (PFR)

The general form of a fuzzy linear regression model can be expressed as follows:

$$\widetilde{Y}_i = \widetilde{A}_0 + \widetilde{A}_1 x_{i1} + \cdots + \widetilde{A}_k x_{ik} = \widetilde{A} x_i \tag{1}$$

where $\widetilde{Y}_i$, $i = 1, 2, \ldots, n$, is the predicted output, which is a fuzzy number; n is the number of data sets; x_{ij}, $j = 0, 1, 2, \ldots, k$ is the jth independent variable of the ith data set; k is the number of independent variables; and $\widetilde{A}_j$ is the fuzzy coefficient of the jth independent variable. $\widetilde{A}_j = \left(s_j^L, a_j^c, s_j^R\right)$, where a_j^c, s_j^L, and s_j^R are the central value, left-, and right-side spreads of the fuzzy coefficients, respectively. If the fuzzy coefficients are symmetric fuzzy numbers, $s_j^L = s_j^R$; $x_i = \left[x_{i0}, x_{i1}, \ldots, x_{ik}\right]$ is a vector of the independent variables and $x_{i0} = 1$, and $\widetilde{A} = \left[\widetilde{A}_0, \widetilde{A}_1, \ldots, \widetilde{A}_k\right]^T$ is a vector of the fuzzy coefficients. The fuzzy regression model, Eq. (1), can be rewritten as follows:

$$\begin{aligned}\widetilde{Y}_i &= \left(\widetilde{Y}_i^{sL}, \widetilde{Y}_i^c, \widetilde{Y}_i^{sR}\right)\\ &= \left(s_0^L, a_0^c, s_0^R\right) + \left(s_1^L, a_1^c, s_1^R\right) x_{i1} + \cdots + \left(s_k^L, a_k^c, s_k^R\right) x_{ik}.\end{aligned} \tag{2}$$

The predicted output of Eq. (1) can be presented as $\widetilde{Y}_i = \left(\widetilde{Y}_i^{sL}, \widetilde{Y}_i^c, \widetilde{Y}_i^{sR}\right)$, where $\widetilde{Y}_i^c$, $\widetilde{Y}_i^{sL}$, and $\widetilde{Y}_i^{sR}$ are the center, left-, and right-side spread values of the output, respectively. The major processes of PFR are described in the following subsections.

2.1. Determination of parameter settings of PDFs

The uncertainty of a random variable can be described by a PDF, $f(x)$, which is a function defined in the interval $[xmin, xmax]$ and has the following properties.

(a) $f(x) \geq 0$ for all x.
(b) $\int_{xmin}^{xmax} f(x)\, dx = 1$.

$xmin$, $xmax$, or both can be infinite.

The form of $f(x)$ depends on the probability distribution of a continuous random variable. Several PDFs, such as uniform, triangular, Gaussian, and exponential functions, are commonly used. The parameter settings of PDFs are determined using COA. COA is a stochastic search algorithm in which chaos is introduced into the optimization strategy to accelerate the optimum seeking operation and determine the global optimal solution (Ren and Zhong, 2011). COA employs chaotic dynamics to solve optimization problems and it has been applied successfully in various areas such as robot optimization control, function optimization and supply chain optimization (Mishra et al., 2008). Compared with conventional optimization methods, COA has faster convergence and can search for better solutions (Nanba et al., 2002). This algorithm also has an improved capacity to seek for the global optimal solution of an optimization problem and can escape from a local minimum. Chaos has dynamic properties, including ergodicity, intrinsic stochastic properties, and sensitive dependence on initial conditions. The characteristic of randomness ensures the capability for a large-scale search. Ergodicity allows COA to traverse all possible states without repetition and overcome the limitations caused by ergodic searching in general random methods. COA uses the carrier wave method to linearly map the selected chaos variables onto the space of optimization variables and then searches for the optimal solutions based on the ergodicity of the chaos variables. The processes of applying COA in this study are described as follows.

First, the number of iterations of COA is defined. Each chaos variable represents the parameter settings of PDFs, and the number of elements in a chaos variable is equal to the number of parameters to be determined. The chaos variable is initialized in which the values are selected randomly in the range [0, 1]. The ranges of parameters $[a, b]$ are initialized, in which a and b are the lower and upper limits of the optimization variable, respectively.

Second, the iteration number is set as $m = 1$. Based on the initialized chaos variable, the logistic model used in COA is shown in Eq. (3), and logistic mapping can generate chaos variables through iteration.

$$c_m = f\left(c_{m-1}\right) = u c_{m-1}\left(1 - c_{m-1}\right), \tag{3}$$

where u is a control parameter; $c_m \in [0, 1]$ is the mth iteration value of the chaos variable c; and c_0 is the initialized chaos variable.

The linear mapping for converting chaos variables into optimization variables is formulated as follows:

$$q_m = a + (b - a) \cdot c_m, \tag{4}$$

where q_m is the optimization variable and the value of q_m is the parameter settings of PDFs. Based on the iteration, the chaos variables traverse between $[0, 1]$, and the corresponding optimization variables traverse in the corresponding range $[a, b]$. In this case, the optimal solution can be identified in the area of feasible solutions.

Based on the values of q_m, PDFs, $f(x)$, are generated. The model can be developed based on $f(x)$ and fuzzy coefficients by which the predicted output $\widetilde{Y}_i = \left(\widetilde{Y}_i^{sL}, \widetilde{Y}_i^c, \widetilde{Y}_i^{sR}\right)$ can be obtained. The predicted crisp output of $\widetilde{Y}_i$ is denoted as $\hat{y}_i$, which is equal to the center value $\widetilde{Y}_i^c$ if symmetric triangular member functions are used in PFR. The mean absolute percentage error ($MAPE$) is defined as the average of percentage errors, which is scale-independent and is a popular measure for evaluating prediction accuracy (Gilliland et al., 2015; Kim and Kim, 2016). Thus, MAPE was adopted in this study as the fitness function in COA, which is defined as follows:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \frac{\left|\hat{y}_i - y_i\right|}{y_i} \cdot 100 \tag{5}$$

where n is the number of data sets; $\hat{y}_i$ is the ith predicted crisp output of $\widetilde{Y}_i$ and y_i is the ith actual crisp output based on survey data. The values