Available online at www.sciencedirect.com
ScienceDirect
Fuzzy Sets and Systems 158 (2007) 2169–2188
FUZZY sets and systems
www.elsevier.com/locate/fss

Multiple regression with fuzzy data

Andrzej Bargiela[a,*], Witold Pedrycz[b], Tomoharu Nakashima[c]

[a]*School of Computer Science, The University of Nottingham, Nottingham NG8 1BB, UK*
[b]*Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alta., Canada T6G 2G6*
[c]*College of Engineering, Osaka Prefecture University, Gakuen-cho 1-1, Sakai, Osaka 599-8531, Japan*

Received 25 July 2005; received in revised form 23 February 2007; accepted 6 April 2007
Available online 20 April 2007

Abstract

In this paper, we propose an iterative algorithm for multiple regression with fuzzy variables. While using the standard least-squares criterion as a performance index, we pose the regression problem as a gradient-descent optimisation. The separation of the evaluation of the gradient and the update of the regression variables makes it possible to avoid undue complication of analytical formulae for multiple regression with fuzzy data. The origins of fuzzy input data are traced back to the fundamental concept of information granulation and an example FCM-based granulation method is proposed and illustrated by some numerical examples. The proposed multiple regression algorithm is applied to one-, three- and nine-dimensional synthetic data sets as well as the 13-dimensional Boston Housing dataset from the machine learning repository. The algorithm's performance is illustrated by the corresponding plots of convergence of regression parameters and the values of the prediction error of the resulting regression model. General comments on the numerical complexity of the proposed algorithm are also provided.

Keywords: Multiple regression; Fuzzy data; Gradient descent; Fuzzy C-means (FCM)

1. Introductory comments

Regression analysis is one of the basic tools of scientific investigation enabling identification of functional relationship between independent and dependent variables. In the classical regression analysis both the independent and dependent variables are given as real numbers. However, in many real-life situations, where the complexity of the physical system dictates adoption of a more general viewpoint, regression variables are given as non-numerical entities such as linguistic variables [6]. Unfortunately, such real-life situations are quite often outside the scope of the classial regression analysis [2,3].

Following the introduction of the concept of fuzzy sets by Zadeh in 1965 [22–24] various researchers attempted extending the regression analysis from crisp to fuzzy domain. The problem statement in this study is diametrically different from the one commonly encountered in the literature. Traditionally, starting from the early work by

* Corresponding author. Tel.: +44 7768634741.
E-mail addresses: Andrzej.Bargiela@cs.nott.ac.uk (A. Bargiela), pedrycz@ece.ualberta.ca (W. Pedrycz), nakashi@cs.osakafu-u.ac.jp (T. Nakashima).

0165-0114/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.fss.2007.04.011

Tanaka et al. [19], see also [18,20,1,16,21], fuzzy regression was introduced as follows:

- Given numeric experimental data $(\mathbf{x}_k, \mathbf{y}_k)$ $k = 1, 2, \ldots, N$, design a fuzzy regression $Y = A_0 \oplus A^{\mathrm{T}}\mathbf{x}$ where A_0 and A are parameters of the model treated as some fuzzy numbers (in particular described by triangular membership functions). Owing to the character of the model and the form of the assumed parameters of the model, Y becomes also a triangular fuzzy number. Note that the operation of addition (denoted here by $\oplus$) pertains to fuzzy numbers rather than plain numeric entities.

In essence, the fuzziness at the output of the regression model emerged because of the lack of perfect fit of numeric data to the assumed linear format of the relationship under consideration. In other words, through the introduction of triangular numbers (parameters), this fuzzy regression reflects the deviations between the data and the linear model. Computationally, the estimation of the fuzzy parameters of the regression is concerned with some problems of linear programming; refer again to the early works in this area. There have been a number of variants of the underlying optimisation techniques, cf. [11,13,16].

In a nutshell, in spite of the differences in the optimisation, the overall mapping can be captured through the relationship

$$\mathbf{R}^n \rightarrow F(\mathbf{R}), \tag{1}$$

where $(F\mathbf{R})$ denotes a family of fuzzy numbers (in our case triangular ones) defined in the space of real numbers $\mathbf{R}$.

The approach advocated in this study marks a departure from the conceptual frameworks governed by (1). For a given collection of input–output fuzzy data, we are concerned in building a numeric regression model that approximates the fuzzy data in an optimal fashion. Referring to (1), the relationship of interest here arises in the form

$$F(\mathbf{R}^n) \rightarrow F(\mathbf{R}). \tag{2}$$

To make the problem of building the regression line more manageable from the optimisation standpoint, we can refine (revise) the mapping to read as

$$F(\mathbf{R}) \times F(\mathbf{R}) \times \cdots \times F(\mathbf{R}) \rightarrow F(\mathbf{R}) \tag{3}$$

with $F(.)$ denoting the corresponding families of fuzzy numbers.

The conceptual framework (2) was originally adopted by Diamond [7,8] who developed a simple linear regression model for triangular fuzzy numbers. This was subsequently generalised to fuzzy regression models with regression variables expressed as arbitrary fuzzy numbers [10,4,5,13,17]. Another generalisation of the regression model, involving the use of fuzzy random variables, was suggested by Koerner and Nather [15]. However, in all of the above approaches the analytical formulae quantifying the values of the parameters of the regression model had to address the issue of negative spreads [9] which complicates significantly the algorithms and makes them difficult to apply to highly-dimensional data. The consequence of having to consider 2^{n-1} (n is dimensionality of data) of optimisation cones in analytical regression methods meant that most of the examples of use of these methods were confined to low-dimensional data, typically single independent and single dependent variable systems.

The main contribution of this paper is the re-formulation of the regression problem as a gradient-descent optimisation, which enables a natural generalisation of the simple regression model to multiple regression models in a computationally feasible way. Indeed, the new formulation provides a basis for a further generalisation to multiple non-linear regression with fuzzy variables.

In Section 2, we provide some background on the classical regression analysis, fuzzy numbers and fuzzy simple linear regression. In Section 3, we extend the scope of fuzzy regression to multiple variables and provide a gradient-descent optimisation algorithm that provides a practical way of calculating regression coefficients. Section 4 offers some practical considerations of generating fuzzy sets for independent and dependent variables and provides several numerical examples of the application of the algorithm to various data sets.

2. Background discussion

2.1. Classical regression analysis

The general task of regression analysis is defined as identification of a functional relationship between the independent variables $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ and dependent variables $\mathbf{y} = [y_1, y_2, \ldots, y_m]$, where n is a number of independent