Information Sciences 180 (2010) 506–518

Contents lists available at ScienceDirect

Information Sciences

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

Modeling manufacturing processes using a genetic programming-based fuzzy regression with detection of outliers

K.Y. Chan [a,b,*], C.K. Kwong [b], T.C. Fogarty [c]

[a] Digital Escosystems and Business Intelligence Institute, Curtin University of Technology, Perth, Australia
[b] Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong
[c] Faculty of Business, Computing and Information Management, London South Bank University, London, UK

ARTICLE INFO

Article history:
Received 13 February 2009
Received in revised form 6 October 2009
Accepted 7 October 2009

Keywords:
Genetic programming
Fuzzy regression
Outlier detection
Epoxy dispensing process

ABSTRACT

Fuzzy regression (FR) been demonstrated as a promising technique for modeling manufacturing processes where availability of data is limited. FR can only yield linear type FR models which have a higher degree of fuzziness, but FR ignores higher order or interaction terms and the influence of outliers, all of which usually exist in the manufacturing process data. Genetic programming (GP), on the other hand, can be used to generate models with higher order and interaction terms but it cannot address the fuzziness of the manufacturing process data. In this paper, genetic programming-based fuzzy regression (GP-FR), which combines the advantages of the two approaches to overcome the deficiencies of the commonly used existing modeling methods, is proposed in order to model manufacturing processes. GP-FR uses GP to generate model structures based on tree representation which can represent interaction and higher order terms of models, and it uses an FR generator based on fuzzy regression to determine outliers in experimental data sets. It determines the contribution and fuzziness of each term in the model by using experimental data excluding the outliers. To evaluate the effectiveness of GP-FR in modeling manufacturing processes, it was used to model a non-linear system and an epoxy dispensing process. The results were compared with those based on two commonly used FR methods, Tanka's FR and Peters' FR. The prediction accuracy of the models developed based on GP-FR was shown to be better than that of models based on the other two FR methods.

1. Introduction

In today's competitive market, manufacturers need to control variability at each of the many processing steps in a manufacturing line, and all variables controlling the desired output in a process need to be understood and optimized to maintain tight control. This can be achieved by developing appropriate physical models to represent the manufacturing process. Physical models [5,9,12,29] are based on a physical understanding of the process, and they typically consist of a set of governing partial differential equations. They are attractive because they provide a fundamental understanding of the relationships between the input and output parameters. However, physical models are usually too complex to be generated accurately for many manufacturing processes.

Statistical regression is a common approach to develop empirical process models [39], but the resulting models are accurate only within the ranges of data from which they are developed. Statistical regression models can be applied only if the

* Corresponding author. Address. Digital Escosystem and Business Intelligence Institute, Curtin University of Technology, Perth, Australia. Tel.: +61 404439097.

E-mail addresses: Kit.Chan@curtin.edu.au (K.Y. Chan), C.K.Kwong@inet.polyu.edu.hk (C.K. Kwong), fogarttc@lsbu.ac.uk (T.C. Fogarty).

0020-0255/$ - see front matter
doi:10.1016/j.ins.2009.10.007

given data is distributed according to a statistical model, and the relationship between dependent and independent variables is crisp. However, in many manufacturing processes, it is difficult to find probability distributions for dependent variables. Artificial neural networks [4,15,20,30,40,45] and fuzzy logic modeling techniques [1,10,18,19,33,47] have been used to develop process models in various manufacturing processes. These approaches normally require a large amount of experimental data to develop models, which are sometimes not available in manufacturing processes. Genetic programming (GP) has been commonly used to develop polynomial models with interaction terms or higher order terms [11,14,24–27,31,32,44,46], but quite a number of manufacturing processes involve uncertainty due to fuzziness that cannot be addressed by GP.

In contrast, a fuzzy linear regression approach in modeling manufacturing processes, which have a high degree of fuzziness, has the distinct advantage of being able to generate models using only a small number of experimental data sets [2,6,21,41–43]. An attempt was made by Schaiable and Lee [38] to model the vertical CVD process using the fuzzy linear regression method. Lai and Chang [28] applied fuzzy linear regression to model the die casting process. Ip et al. [16] used fuzzy linear regression to develop a process model for epoxy dispensing. Modeling of transfer molding using fuzzy linear regression was also reported by Ip et al. [17]. Kwong and Bai [22] performed process modeling and optimization using both fuzzy linear regression and fuzzy linear programming approaches. Three different approaches of fuzzy linear regression were summarized in Chang and Ayyub [3]. However, existing fuzzy regression (FR) approaches cannot be used to develop models that contain interaction terms or higher order terms. In fact, behavior of many manufacturing processes is non-linear. If interaction terms or higher order terms could be considered in FR, models which provide more accurate prediction of manufacturing processes would be developed. Furthermore, it is widely recognized that the quality of model development declines when outliers in experimental data exist, but very few studies have attempted to detect outliers when developing FR models. Chen [8] proposed a method to detect outliers involving crisp inputs and fuzzy outputs. The method detects the difference in width between the spread of fuzzy data and the spread of fuzzy output. However, experimental data and the results of manufacturing processes involve crisp values of experimental settings and crisp values of experimental responses. Therefore, the method cannot be applied to manufacturing processes.

These modeling methods ignore both the interaction terms (or higher order terms) in manufacturing processes as well as the fuzzy nature of data. Moreover, they produce black-box models not usually recommended by process engineers, and they include outliers in model development or require a large amount of data to produce models, that are usually not available in real situations. These modeling methods cannot address the entire range of characteristics of the manufacturing process. To overcome thes‘e deficiencies, we propose genetic programming-based fuzzy regression (GP-FR), which can be used to generate models with interaction or higher order terms. GP-FR uses the general outcomes of GP to construct models based on a tree structure representation in which both the interaction and higher order terms can be considered. The FR generator is also proposed to detect the outliers from experimental data sets based on an indicator of outliers. The FR generator then estimates the contribution of each branch of the tree in order to determine the fuzzy coefficient of each term of the model by using the experimental data sets excluding the outliers. As interaction and higher order terms can be generated and represented in the branches of the tree based on the GP-FR approach, FR models in fuzzy polynomial form with interaction and higher order terms can be generated as explicit models. Furthermore, as the FR generator is used to determine fuzzy coefficients of the model, only a small amount of data is required to generate the process models, which is practical in the manufacturing process.

The effectiveness of the proposed GP-FR approach is evaluated by modeling simple non-linear systems and the epoxy dispensing process for electronic packages, which is used in various electronic packaging processes such as integrated circuit (IC) encapsulation, die-bonding, and placement of surface mount components [7]. In today's competitive market, the process parameters of the epoxy dispensing process, which directly affects the quality of electronic packaging products, need to be understood and optimized. However, epoxy dispensing is a highly non-linear process that involves extremely complex interrelationships among the epoxy properties, process conditions and overall encapsulation quality [13]. GP-FR is used to develop models for this manufacturing process. Modeling results based on GP-FR is compared with those based on the fuzzy linear regression methods of Tanaka [43] and Peters [36], which have been employed to model the epoxy dispensing processes.

2. Fuzzy regression

The FR model can be developed based on M experimental data sets $\{(y_1, \boldsymbol{x}_1), (y_2, \boldsymbol{x}_2),\ldots,(y_i,\boldsymbol{x}_i),\ldots,(y_M,\boldsymbol{x}_M)\}$. $\boldsymbol{x}_i$ is the i th experimental data set of the explanatory variable, $\boldsymbol{x}_i = (x_{i0},x_{i1},\ldots,x_{ij},\ldots,x_{iN})$, where $x_{i0} = 1$ for all i, and x_{ij} is the observed value of the jth variable in the ith experimental data set and is always crisp. y_i is the ith observation of the explained variable, $i = 1,2,\ldots,M$, and it is a crisp value. In particular, the fuzzy linear regression model can be represented as follows:

$$\tilde{y} = \tilde{f}(\boldsymbol{x}) = \widetilde{A}_0 + \widetilde{A}_1 x_1 + \cdots \widetilde{A}_j x_j \cdots + \widetilde{A}_N x_N \tag{1}$$

where $\tilde{y}$ is the estimated observation after adjusting $\widetilde{A}_0, \widetilde{A}_1, \ldots, \widetilde{A}_N$. In FR models, the disturbance is not introduced as a random addend in the linear relation, but it is incorporated into the fuzzy coefficients $\widetilde{A}_j (j = 0, 1, \ldots, N)$. The FR problem is to determine the fuzzy coefficients $\widetilde{A}_j = \left(a_j^c, a_j^s\right) (j = 1, 2, \ldots, N)$ with the central point a_j^c and the spread a_j^s of $\widetilde{A}_j$ such that the total systematic fuzziness is minimized, while the given input–output pairs should be included in their h-level set described as $y \in [\tilde{y}]_h$ [43]. It can be formulated as a linear programming (LP) problem as follows: