Information Fusion 27 (2016) 161–169

Contents lists available at ScienceDirect

Information Fusion

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

Outlier elimination using granular box regression

M. Reza Mashinchi [a], Ali Selamat [a,b,*], Suhaimi Ibrahim [c], Hamido Fujita [d]

[a] *Faculty of Computing, Universiti Teknologi Malaysia, Johor, Malaysia*
[b] *UTM-IRDA Digital Media Center of Excellence, Universiti Teknologi Malaysia, Johor, Malaysia*
[c] *Advance Informatics School, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia*
[d] *Intelligent Software Laboratory, Iwate Prefectural University (IPU), Iwate, Japan*

ARTICLE INFO

Article history:
Received 1 October 2014
Received in revised form 8 April 2015
Accepted 12 April 2015
Available online 20 April 2015

Keywords:
Granular box regression
Outlier elimination
Noisy data
Data simplification
Data abstraction

ABSTRACT

A regression method desires to fit the curve on a data set irrespective of outliers. This paper modifies the granular box regression approaches to deal with data sets with outliers. Each approach incorporates a three-stage procedure includes granular box configuration, outlier elimination, and linear regression analysis. The first stage investigates two objective functions each applies different penalty schemes on boxes or instances. The second stage investigates two methods of outlier elimination to, then, perform the linear regression in the third stage. The performance of the proposed granular box regressions are investigated in terms of: volume of boxes, insensitivity of boxes to outliers, elapsed time for box configuration, and error of regression. The proposed approach offers a better linear model, with smaller error, on the given data sets containing varieties of outlier rates. The investigation shows the superiority of applying penalty scheme on instances.

1. Introduction

Simplifying and abstracting comfort us to understand the data to trace its general pattern or trend. While the term of “abstracting” associates with studies in artificial intelligence, the term of “granularity” is its synonym in soft computing studies [1–4]. Granularity often aims at reducing the complexities of data that cause to increase the processing cost – mostly where the uncertainties are involved. Therefore, the certain practices motivate the studies on granularity, i.e., clarity, low cost approximation and the tolerance of uncertainty. An application of these practices, through understanding the data, are identifying or eliminating the anomalies – known as outliers. The granulated data can benefit the computation in two ways as follows. (i) To understand the data: by transparent the complexity to a reduced size data – known as granules. When a method performs an estimation based on granular data, as the requisite, the outcome should be more accurate than using the original complex data. (ii) To reduce the cost of data analysis: by avoiding complex tools to run through the data for insight discovery [2,5–7]. Thus, a non-expert data mining person can also make sense out of it. (iii) To increase the power of estimation and the capability of dealing with uncertainty. A data, as a part of a granule, does not only represent a single observation but, rather, a group of data.

However, beside the pointed advantages for granulation, it requires methods to sharpen the transparency of data. Granular box regression analysis [8,9] carries this out by detecting the outliers in data. It finds the correlation between the dependent and independent variables using hyper dimensional interval numbers known as boxes. In granular box regression (GBR), every instance in data set affects the size and coordinate of boxes. In the result, an approach becomes sensitive to outliers as an issue. To resolve this, we propose variations of granular box regression based on subset of data they process, and then, we investigate their performance with the presence of outliers in a data set. There are two motivations to the proposed variations. First, to simplify a data set contains numerous data – thus, we comfort understanding of a non-expert by clarifying the relationship between dependent and independent variables; second, to study the performances of each variation of granular box regression with presence of outliers.

Outlier is an anomaly object that is atypical to the remaining data. It deviates from other objects such that it makes suspicious to be generated by a different mechanism [10]. Subject of outlier is either an elimination of disturbance such as noise reduction [10–14], or an interest of detection such as crime detection [15–23]. This paper focuses on the former view. Different approaches [10,11,15,24] have studied the outlier; though, this paper differentiates itself by applying granular box approach and elimination of

[*] Corresponding author at: Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia.

http://dx.doi.org/10.1016/j.inffus.2015.04.001
1566-2535/© 2015 Elsevier B.V. All rights reserved.

the outliers. To measure the goodness of applied approach, either box or relationships between boxes can indicate the quality of box configuration on data. In case of box measurement, the approach should minimize the overall volumes to reduce the complexity of data as were indented; and in case of measuring the relationships of boxes, am approach should build a coherent relationship similar to the true function by regression analysis. A possible approach, to furnish the former issue, is employing genetic algorithm (GA) [25–27] to find the optimal volumes gradually. Performing the box configuration based on GA builds the relationships between boxes to reduce the complexity of data and exclude the outliers. As the result, configured boxes represent the simplified original data.

This paper is organized in seven sections. Section 2 reviews the granular box regression and explains its notions. Section 3 explains the three stages of proposed framework to configure the boxes, eliminate the outliers, and fit the curve; where, box-based penalization (BP) and instance-based penalization (IP) are proposed for box configuration, and clean- and candidate-based methods are proposed for elimination of outliers. Section 5, reveals the results on six data sets each with two rates of outliers. Then, Section 6 gives detailed analyses in three parts to investigate the regression analysis and box configuration with respect to effect of dimensionality of data and rate of outliers on each method. Section 7 concludes the achieved results and addresses the future works for each method.

2. Granular box regression

A regression approach should eliminate the outliers prior to fit a model on data; as otherwise, it may fit the outlier data and makes a wrong interpretation. Granular box regression (GBR) is an inclusive approach to detect the outliers in every dimension of data. Compared with classical regression analysis (CRA), which performs only on respond variable [27–29], GBR performs also on predictors to detect the outliers. Where CRA approach minimizes the summation of distances between the actual and the estimated values, the GBR approach obtains the minimum summation of volumes of all boxes. In general, GBR is a granular regression approach involved with fuzzy granulation generalization, f.g-generalization, to find the relationships of boxes [2,30–33]. Fig. 1 shows fuzzy graph representation of f.g-generalization.

GBR addresses three methods [9] of borderline, residual, and average distance. Borderline method examines the significant decrease of box volume by eliminating a potential outlier on box borders. Residual method examines the significant largeness of distance between the residual errors of a potential outlier and the average of all data, where the residual error is a distance between estimated and true value. Average distance method examines the significant largeness of distance between an individual distance of a potential outlier and the aggregated distance, where an individual distance is the average distance of a data to all others within a box and the aggregated distance is the average distance of all data to each other within a box [8]. Eqs. (1) and (2) defines the individual and aggregated distances; where for *Box*: *All* is the number of data, D_i and D_j are the ith and jth instances, and, k_{Box} is the kth instance in a *Box*.

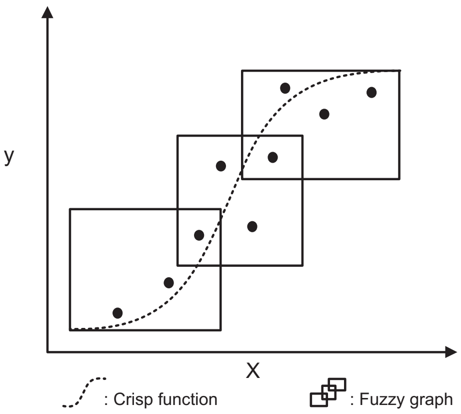

Fig. 1. Fuzzy graph representation of f.g-generalization.

$$Distance_{k_{Box}}^{individual} = \frac{\sum_{i=1}^{All-1} \|D_i - D_{k_{Box}}\|}{All - 1} \tag{1}$$

$$Distance_{Box}^{agregated} = \frac{\sum_{i=1}^{All} \sum_{j=i+1}^{All} \|D_i - D_j\|}{\sum_{i=1}^{All-1} (All - 1)} \tag{2}$$

Earlier approach, P, for GBR [8] examines the box volume based on borderline method to identify the actual outliers out of potential outliers. In contrast, this paper proposes based on average distance method. In both approaches, P and our proposed, the mathematical formulation to obtain the optimized coordinates of n boxes for an m-dimensional problem is as follows:

$$Minimize \left(\sum_{k=1}^{n} V(B^k) \right) \tag{3}$$

where V calculates the volume of a given box such as B^k. Note that B^k is kth box that four vertices can represent it in a two dimensional space. Both approaches initiate the number of boxes prior to perform the box regression analysis. The larger number of boxes the higher resolution of regression data; conversely, the less number of boxes the more abstract derived relationships. Studying the optimized number of boxes is in the scope of future works, tough it significantly affects on GBR.

3. Proposed granular box regression

An overall view to our proposed GBR is shown in Fig. 2. It illustrates the idea of penalty scheme, where box configuration represents the simplification of data by clean instances or candidates of boxes. Concretely, Fig. 3 shows the procedure of performing GBR in three stages that are: (i) apply a granular box configuration, (ii)

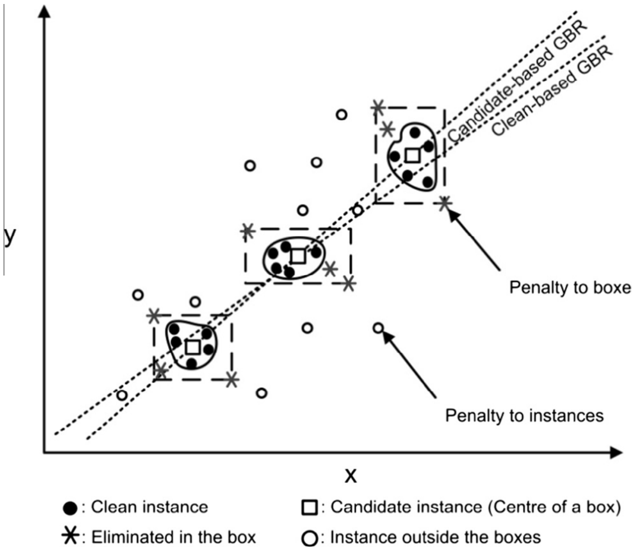

Fig. 2. An overall view to the proposed penalty scheme GBRs.