Robust Algorithms for Regression Analysis Based on Fuzzy Objective Functions

Tai-Ning Yang*, Chih-Jen Lee, Jenn-Dong Sun, Chun-Jung Chen

Department of Computer Science and Information Engineering, Chinese Culture University, Taipei, Taiwan, ROC.

Received 11 July 2017; received in revised form 17 August 2017; accepted 26 August 2017

Abstract

In this paper, we address the issues related to the design of fuzzy robust linear regression algorithms. The design of robust linear regression analysis has been studied in the literature of statistics for over two decades. More recently various robust regression models have been proposed for processing noisy data. We proposed a new objective function by using fuzzy complement and derive improved algorithms that can produce good regression analysis from the spoiled data set. Data set from the U.S. Department of Transportation is used to evaluate the performance of the regression algorithms.

Keywords: robust regression, fuzzy complement, linear regression analysis

1. Introduction

Linear regression analysis is the study of linear relationships between variables. As a basic and 79 popular statistical technique, it has been widely used in various fields. Since linear regression analysis algorithms have to process data from the real world, it should have the ability to cope with the outlier defined as the observation point that is distant from other observ ations. Robustness theory is concerned about solving problems subject to model perturbation or outlier. According to Huber [1], a robust algorithm not only performs well under the assumed model, but also produces a satisfactory result under the deviation of the assumed model.

More recently, many researchers proposed various robust algorithms for regression analysis. Kopsinis et al. [2] proposed a mechanism for iteratively detecting and excluding corrupted data. Papageorgiou et al. [3] splited the noise into two compon ents: the inlier bounded noise and the outliers. They constructed a robust method in the framework of greedy algorithms. Huang et al. [4] developed an effective convexapproach that used recent advances on rank minimization and applied the method in computer vision applications. Cheng et al. [5] introduced a robust adaptive loss function to measure the representation loss. Nurunnabi et al. [6] used global polynomial functions and designed robust algorithms for extracting the ground points in laser scanning 3-D point cloud data. Unlike previous approaches, our proposed robust regression analysis is based on fuzzy objective functions.

2. Traditional Linear Regression Analysis

Regression analysis is a statistical process for estimating the dependent variable from one or more independent variables. The target dependent variable is formulated as a function of the independent variables called the regression function. When the function is linear, the process is called linear regression analysis.

* Corresponding author. E-mail address:tnyang@faculty.pccu.edu.tw

$D_i = < x_i, y_i >$ denotes the i-th data pair. $x_i = (x_i^1, x_i^2, .., x_i^k)$ is the vector of k independent variables and y_i is the target dependent variable. There are n data pairs.

$f(x_i) = w_0 + w_1 x_i^1 + w_2 x_i^2 + .. + w_k x_i^k$ is the linear estimation function that is the linear combination of input components. The weight $w = (w_0, w_1, .., w_k)$ is the coefficients vector for estimation. $e(x_i) = (y_i - f(x_i))^2$ is the loss function.

The objective function of traditional linear regression analysis for minimization is $\sum_{i=1}^{n} e_i$.

The following is the online algorithm of linear regression derived by gradient descent approach.

Step 1. Initially set the iteration count *t* , iteration bound *T*, learning coefficient $\alpha_0 \in (0,1]$ and the weight *w*.

Step 2. While *t* is less than *T*, do steps 3-7.

Step 3.Compute $\alpha_t = \alpha_0 (1 - t / T)$ and set $i = 1$

Step 4. While *i* is less than *n*, do steps 5-6.

Step 5. Update the weight:

$$w_0^{new} = w_0^{old} + \alpha_t (y_i - f(x_i)) \tag{1}$$

$$w_1^{new} = w_1^{old} + \alpha_t (y_i - f(x_i)) x_i^1 \tag{2}$$

$$w_k^{new} = w_k^{old} + \alpha_t (y_i - f(x_i)) x_i^k \tag{3}$$

Step6. Add 1 to *i*.

Step 7. Add 1 to *t*.

The above algorithm is known to fail when outliers exist.

3. Robust Linear Regression Analysis Based On Fuzzy Objective Functions

For tackling the noise, we add a noise cluster in which the data has a constant influence η . Assume that there is an outlier cluster outside the data cluster. u_i is the membership of x_i in the data cluster, while the standard fuzzy complement of u_i , $(1 - u_i)$, is the membership of x_i in the noise cluster. The fuzziness variable, *m*, determines the influence of small u_i compared to large u_i . Following the fuzzy theory, we propose a robust linear regression objective function:

$$RLG = \sum_{i=1}^{n} (u_i)^m e(x_i) + \eta \sum_{i=1}^{n} (1 - u_i)^m , \tag{4}$$

Subject to u_i in [0,1] and *m* in [1, ∞).

Compute the gradient of RLG with respect to u_i and get

$$u_i = \frac{1}{1 + (\frac{e(x_i)}{\eta})^{1/(m-1)}} , \text{ when } RLG \text{ is minimized.} \tag{5}$$

Substituting this membership back and after simplification, we get

$$RLG = \sum_{i=1}^{n} \left(\frac{1}{1+(\frac{e(x_i)}{\eta})^{1/(m-1)}} \right)^{m-1} e(x_i) \tag{6}$$

Following the multidimensional chain rule, the gradient of *RLG* with respect to *w* is:

$$\frac{\partial RLG}{\partial w} = (\frac{\partial RLG}{\partial e(x_i)})(\frac{\partial e(x_i)}{\partial w}) = (\frac{1}{1+(\frac{e(x_i)}{\eta})^{1/(m-1)}})^m (\frac{\partial e(x_i)}{\partial w}) \tag{7}$$

Let $\beta(x_i)$ denote $(\frac{1}{1+(\frac{e(x_i)}{\eta})^{1/(m-1)}})^m$. *m* is called the fuzziness variable in the literature of fuzzy clustering. The following is the proposed algorithm.

Step 1. Initially set the iteration count *t*, iteration bound *T*, learning coefficient $\alpha_0 \in (0,1]$, soft threshold η and the weight *w*.

Step 2. While *t* is less than *T*, do steps 3-7.

Step 3.Compute $\alpha_t = \alpha_0(1-t/T)$, set $i=1$.

Step 4. While *i* is less than *n*, do steps 5-6.

Step 5. Update the weight:

$$w_0^{new} = w_0^{old} + \alpha_t \beta(x_i)(y_i - f(x_i)) \tag{8}$$

$$w_1^{new} = w_1^{old} + \alpha_t \beta(x_i)(y_i - f(x_i))x_i^1 \tag{9}$$

$$w_k^{new} = w_k^{old} + \alpha_t \beta(x_i)(y_i - f(x_i))x_i^k \tag{10}$$

Step 6. Add 1 to *i*.

Step 7. Add 1 to *t*.

4. Simulations

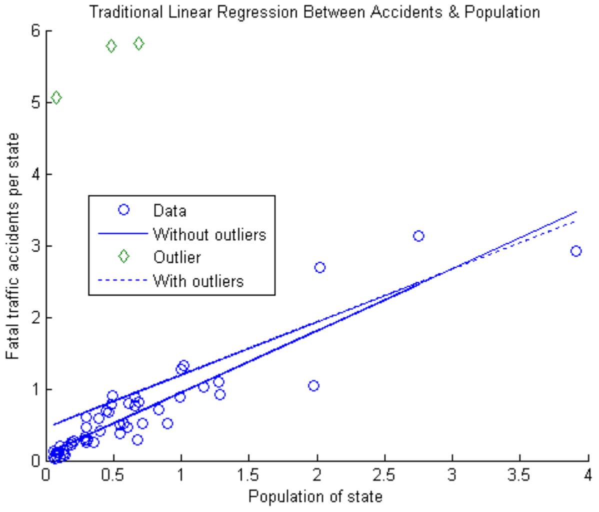

Fig. 1 Results of traditional linear regression

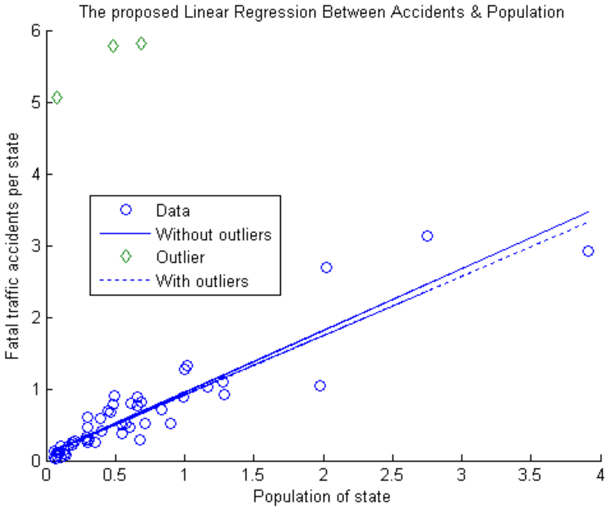

Fig. 2 Results of the proposed linear regression

To show the experimental differences between the traditional and the proposed, we use the data set from the U.S. Department of Transportation. The data pair is the population and fatal motor vehicle crashes per state in 2015. There are 51 pairs corresponding to 50 states and the District of Columbia. For convenience, we scale down the data to be (population/107, crashes/103). In the following experiments, we set iteration bound T =1000, learning coefficient $\alpha_0 = 0.3$ and the fuzziness variable $m = 3$. The noisy data set is generated by adding 5000 crashes to the first 3 data. Fig. 1 shows the traditional linear regression is greatly affected by the outliers while Fig. 2 shows the proposed one is slightly affected by the outliers. As suggested by Huber [1], the constant influence η is set as the mean of the $e(x_i)$ from the result of the traditional linear regression. The initial weight *w* in the robust approach is also set by the result of the traditional linear regression.

5. Conclusions

With consideration of outliers, we propose a fuzzy objective function for robust linear regression. The derived algorithm adapts the estimated component according to the current membership of the input data. Thus, the influence of outliers is alleviated. The results of a simple simulation comparing the traditional linear regression and the robust linear regression correspond to our expectations.

Acknowledgement

The support of the Ministry of Science and Technology, (Taiwan), under Grant Most 105-2221-E-034-016 is gratefully acknowledged.

References

[1] P. J. Huber and E. M. Ronchetti, Robust statistics, 2nd, New York: Wiley, 2009.

[2] Y. Kopsinis, S. Chouvardas, and S. Theodoridis, “Iterative randomized robust linear regression,” International Conference on Acoustics, Speech and Signal Processing, pp. 5436-5540, April 2015.

[3] G. Papageorgiou, P. Bouboulis, and S. Theodoridis, “Robust linear regression analysis—a greedy approach,” IEEE Transactions on Signal Processing, vol. 63, no. 15, pp. 3872-3887, May 2015.

[4] D. Huang, R. Cabral, and F. De la Torre, “Robust regression,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 2, pp. 363-375, February 2016.

[5] G. L. Cheng, F. Y. Zhu, S. M. Xiang, Y. Wang, and C. H. Pan, “Semisupervised hyperspectral image classification via discriminant analysis and robust regression,” IEEE Journal of selected topics in applied earth observations and remote sensing, vol. 9, no. 2, pp. 595-608, September 2016.

[6] A. Nurunnabi, G. West, and D. Belton, “Robust locally weighted regression techniques for ground surface points filtering in mobile laser scanning three dimensional point cloud data,” IEEE Transaction on Geoscience and Remote Sensing, vol. 54, no. 4, pp. 2181-2193, November 2015.