Journal of Education and Science
Vol. 27, No.3, 2018, pp. 136-142
ISSN 1812-125X
http://edusj.mosuljournals.com

Using Genetic Algorithm in Outlier Detection for Regression Model

Zakariya Y. Algamal* Hamsa M.Thabet**

*Dept. of Statistics and Informatics, college of Computers Sciences and Mathematics. **Dept. of Information and libraries, college of Arts.

تاريخ الاستلام 2012/10/07 *تاريخ القبول 2013/01/09*

الخلاصة

يعتبر تحليل الانحدار الخطي من أكثر الأساليب الإحصائية استخداما في تحليل البيانات في اغلب التطبيقات. في بعض الأحيان تحتوي البيانات قيد البحث على مجموعة من القيم الشاذة ويكون من الضروري جدا تشخيص هذه القيم لضمان صحة التحليل الإحصائي. في هذا البحث استخدمنا الخوارزمية الجينية مع ثلاث أنواع من دوال الهدف وهي معيار اكاكي للمعلومات , معيار بيز للمعلومات , ومعيار هانان – كيون للمعلومات لتشخيص مشكلة التقنع والإخفاء للقيم الشاذة في نموذج الانحدار الخطي . تم استخدام مجموعتين من البيانات المدروسة مسبقا والمعتمدة عالميا في بحثنا هذا تم التوصل الى ان استخدام الخوارزمية الجينية في تشخيص القيم المقنعة والمخفية مقارنة باستخدام معيار أكاكي ومعيار هافانركون للمعلومات كدوال للهدف مقارنة بمعيار بيز للمعلومات .

Abstract

Linear regression model is commonly used to analyze data from many fields. Sometimes the data under research contains outliers, and it is important that these outliers be identified in the course of the correct statistical analysis. In this article we used genetic algorithm (GA) with three type of objective functions,Akaike information criterion (AIC), Bayesian information criterion (BIC), and Hannan–Quinn information criterion (HQIC) to detect the problem of masking and swamping outliers in linear regression model . Two well – known data sets have been studied and we conclude that GA doing-well in detection these type of outliers when using AIC and HQIC comparingwithBIC.

Keywords: *outliers king ; swamping ; Genetic Algorithm ;information criteria*

Introduction

Linear regression model is the most widely used approach for fitting models to data . Identifying regression outliers can be very important task in many fields of science . They should be identified since of their potential effect on parameter estimates and inference form the regression model . An outlier is an observation or a group of observations which different form the bulk of the data . If a regression data set contains only a single outlier , many procedures have been proposed in the past years . However when a group of observations clustered , a standard detection procedures like studentized residuals , DFFITS , and Cook's distance would fail to identify the outliers (masking) or can identify the inlaying observations as outliers (swamping) . Although , of these two types of errers , masking can be viewed as a more serious problem than swamping [10] One of the resent inntellegent approach used in detecting masking and swamping is genetic algorithm (GA). Tolvi [11] used Bayesian information criterion as a fitness function , while Alma and et al [2] used GA in detecting outlier when the multicollinearity problem present in linear regression model . They used the same objective functions as Tolvi . In this paper we use different fitness functions, Akaike information and Hannan- Quinn information criteria.

Outlier in Linear Regression

We will consider the following classical linear regression mode:

$$Y = X\beta + \varepsilon \dots\dots\dots (1)$$

Where Y is an (n×1) vector of the response variable and X is an (n×p) matrix representing p explanatory variables, β is an (p×1) vector of unknown n parameters , and Ɛ is an (n×1) vector of error whose conditional mean and variance are 0 and σ^2In respectively .Using least squenes method in estimating β in equation (1) , where :

$$\hat{\beta} = (x'x)^{-1}\, x'y \dots\dots\dots (2)$$

Observations that do not follow the same model as the rest of the data are called outliers , many methods have been suggested for detecting single outliers such as studentized residuals, DFFITS, and Cook's distance.

We delete the i th observation and use the remaining (n-1) observations to calculate the fitted value of the i th case , $\hat{Y}i(i)$ [3] .

Masking and Swamping effects in Linear Regression

The masking effects means that an outliers is undetected because of the presence of another ones , and swamping effects is that a good observation is incorrectly identified as an outlier because of presence of outlier clean subset . Both masking and swamping are produced from the OLS method together with the mean –shift outliers models [4]. In other word, it is said that one outlier masks a second outlier , if the second outlier can be considered as an outlier only by itself , but not in the presence of the first outlier . thus , after the deletion of the first outlier the second observation is emerged as an outlier . it is said that one outlier swamps a second observation , if the latter can be considered as an outlier only under the presence of the first one .

Genetic Algorithm and Swamping Detection

Genetic algorithm (GA) is a stochastic search algorithm based on the mechanism of natural selection and natural genetics [7] GA , in contrast to classical search techniques , start with an initial set of random solutions called population . each individual in the population is called a chromosome , encoding a solution to the problem at hand . A chromosome is a string of symbols , usually but not necessarily , a binary bit string. The chromosomes evolve through successive iterations , called generations. During each generation, the chromosomes are evaluated, using some measures of fitness. To Create the next generation, new chromosomes which called offspring are formed by either merging two chromosomes from the current generation using a crossover, mutation and selection operators. After several generations, the algorithm converges to the best chromosome which we hope represents the optimal solution to the problem when decoded [8].

The etection of outliers is important because the inference drawn from the model will be biased if outliers are not treatment. In our paper we use GA as a tool to detect the masking and swamping outliers that may be affect the regression results which the classical approach may have long time to detect them. We will start by describing the coding of the candidate models for outlier detection, each model will be a binary vector:

$$w = (w1, w2, \dots wn), where :$$

$$wi = \begin{cases} 1 & if\ wi\ indicates\ an\ outlier \\ 0 & otherwise \end{cases} \quad \dots\dots\dots(3) \qquad :$$

The measure of fitness of a chromosome is evaluated by using three objective functions , Bayesian information criterion (BIC) [9], Akaike

information criterion (AIC) [1] , and Hannan [6] –Quiun information criterion (HQIC) [3], where:

$$BIC = n[\log(1-R2)+(1-K)\log(n)+md\log(n)].....(4)$$

$$AIC = n[\log(1-R2)+m].....(5)$$

$$HQIC = n[\log(1-R2)+\log(\log(n))].....(6)$$

Where R^2 is the coefficient of the determination and k is the number of kappa explanatory variables, (γ) is the extra penalty given to outliers dummies , nd is the number of outlier dummies , and m = 1+k+nd. The minimum value of the three objective functions will be chosen which represent the diagnostic of the real outliers.

5-Examples

In this paper, we use Two well-known data sets to illustrate the masking and swamping outliers detection in linear regression model . In our algorithm, the population size in each generation is 100, selection is of type remainder, mutation was uniform , and the crossover was tow point type. Two examples was taken that have been diagnostic to have masking and swamping effects.

(5-1) Example (1)

In this example we introduce the Hadi and Simonoff artificial data [5]. They assume two explanatory variables x1 and x2 which they distributed as uniform (0,15) and use the regression model $y = x1 + x2 + \varepsilon$, where ε has normal distribution with mean zero and standard deviation 1 . The first three observation is assumed to have a mean-shift outlier with $y = x1 + x2 + 4 + \varepsilon$. Table (1) shows the minimum objective values GA results.

Table (1) :The GA results of Example (1)

Observation detected	AIC	BIC	HQIC	Generation numbers	The number of the generated solution
1	8.339	10.148	8.552	50	50
2	9.211	10.161	9.3801	75	51
3	9.151	10.273	9.023	75	51
1,2	8.412	10.107	8.143	75	70
1 ,3	7.818	10.075	8.003	80	80
2 , 3	8.033	10.092	7.74	85	85
1 , 2 , 3	6.973	9.913	7.589	85	85

As we see from table (1) that GA doing well detecting the masking outliers comparing with the classical detection procedures which they failed in detecting. All three objective functions AIC, BIC and QHIC succeed in detection the observations.

(5-2) Example (2)

Here we used that data given by Atkinson and Riani [Atkinson & Riani , 2000] . The data consist of 60 observation on three explanatory variables where there are six masked outliers that cannot be detected using traditional procedures . The observations are 2, 9 , 30, 31 ,38 ,47 and 21 , as well as the observation that detected by traditional method is 43 .Table (2) shows the results.

Table (2) : Masking detection in Example(2)

Observation detected	AIC	BIC	HQIC	Generation numbers	The number of the generated solution
43	10.94	24.222	20.0135	80	80
2,9,30,31,38,43,47	9.864	19.873	19.873	85	85

From table (2) , Once conclude that the traditional method detect the observation 43 as an outliers but , they failed in detection the masking outliers 2,9,30,31,38,43 and 47 comparing with GA that be better in detection outliers.

Conclusions

1- Genetic algorithm doing well in detection masking outliers as seen from the two examples or swamping comparing with the traditional methods such as Cook's distance and DFFITS ,which these methods depending on single observation deletion .

2- Genetic algorithm save the time of computation unlike classical procedures , which takes several minutes , where in example (1) detection masking outliers in GA takes 10 second . The same time in example (2) , where it takes more than 15 minutes in classical procedures at least.

3-The three objective functions agreed to the same masking outliers detection , although all papers that consider the problem of detecting outliers using GA depending on BIC only.

4- AIC gave less values as a objective function comparing with the other two objective functions, BIC and HQIC. Where BIC gave largest values comparing with AIC and HQIC, thus we conclude and support to use AIC and HQIC as objective function.

References

1-Akaike , H. , 1974, " A new look at the statistical model identification " , IEEE transactions on Automatic control vol.19, No. 6 , pp.716 – 723.

2- Alma , O. , G. , Kurt , S. and Uour , A. , 2009 , " Genetic Algorithm Based outlier Detection Using Bayesian Information Criterion in Multiple Regression Models Having Multicollinearity problems " , Gazy University , Journal of science vol.22 , No. 3, pp. 141 – 148 .

3- Atkinson , A. and Riani , M. , 2000," Robust Diagnostic Regression Analysis " , Springer – Verlag , New York .

4- Chiang , J. , 2008 , " The Algorithm for multiple outliers Detection against Masking and Swamping Effects " , International of contempt Mathematical Sciences , vol.3 , No. 17, pp. 839 – 859 .

5- Hadi, A.,S. and Simonoff , J.,S., 1993 , " Procedures for the Identification of Multiple outliers in Linear Models " , Journal of the American Statistical Association , vol.88 , No.424 , pp.1264 – 1272.

6- Hannan , E.,J. and Quinn , B.,G.,1979, " The determination of the order of an Auto regression ", Journal of Royal Statistical Society , B, vol.41, pp.190 –195 .

7- Raja , P. , V. and Bhaskaean , V.,M., 2012," An Effective Genetic Algorithm for outlier Detection " , International Journal of Computer Applications , vol.38 , No. 6, pp. 30 – 33 .

8- Sivanandam and Deepa , 2008," Introduction to Genetic Algorithm , Springer – Verlag Berlin Heidelberg , New York .

9- Schwarz , G.,E., 1978 , " Estimating the dimension of a model " , Annals of statistics , vol.6 , No.2, pp.461-464.

10- Siniksaran , E. and Satman ; M. , 2011,"PURO :A package for Unmasking Regression outliers", Gazi university Journal of science vol.24,No.1,pp. 59– 68 .

11- Tolvi , T. , 2004 , " Genetic Algorithms for outlier detection and variable selection in Linear Regression Models " , soft computing , vol.8 , 2008 , pp.527 – 533.