Outlier Detection in Logistic Regression: A Quest for Reliable Knowledge from Predictive Modeling and Classification

Abdul Nurunnabi[*], Geoff West[†]
Department of Spatial Sciences
Curtin University, Perth, Australia
Cooperative Research Centre for Spatial Information (CRCSI)
[*]abdul.nurunnabi@postgrad.curtin.edu.au, [†]g.west@curtin.edu.au

Abstract— **Logistic regression is well known to the data mining research community as a tool for modeling and classification. The presence of outliers is an unavoidable phenomenon in data analysis. Detection of outliers is important to increase the accuracy of the required estimates and for reliable knowledge discovery from the underlying databases. Most of the existing outlier detection methods in regression analysis are based on the single case deletion approach that is inefficient in the presence of multiple outliers because of the well known masking and swamping effects. To avoid these effects the multiple case deletion approach has been introduced. We propose a group deletion approach based diagnostic measure for identifying multiple influential observations in logistic regression. At the same time we introduce a plotting technique that can classify data into outliers, high leverage points, as well as influential and regular observations. This paper has two objectives. First, it investigates the problems of outlier detection in logistic regression, proposes a new method that can find multiple influential observations, and classifies the types of outlier. Secondly, it shows the necessity for proper identification of outliers and influential observations as a prelude for reliable knowledge discovery from modeling and classification via logistic regression. We demonstrate the efficiency of our method, compare the performance with the existing popular diagnostic methods, and explore the necessity of outlier detection for reliability and robustness in modeling and classification by using real datasets.**

Keywords - data mining; high leverge point; influential observation; knowledge discovery; outlier; pattern recognition; regression; reliability; statistical computing

I. INTRODUCTION

It has been recognized that the Knowledge Discovery in Databases (KDD) community ignored statistical methods on the basis of courses that they took many years ago [1]. Logistic regression (LR) is a statistical technique mainly used for modeling for the case of a categorical (binomial and multinomial) response variable. From its original domain in epidemiology and health research, the applications of LR have expanded to cover almost every branch of knowledge including business, marketing, engineering, criminology, ecology, space and spatial sciences [2, 3]. In recent years, it has drawn huge attention as a successful Data Mining (DM) and high dimensional classification tool [4]. DM is an intermediate step in KDD (Fig. 1) involving algorithms that explore databases, extract patterns and develop models for analysis and prediction. The success of DM depends heavily on the algorithms and techniques used. Before performing DM techniques one of the steps of the KDD paradigm (Fig. 1) [5, 6] is data processing, which covers data cleaning and preparing data to ensure accurate results, making proper decision and helping to obtain reliable knowledge. It is known that the presence of outliers can make a technique unreliable, give inaccurate, non-robust results, and draw imperfect and erroneous inferences. Knorr et al. [7] identify outlier detection as a meaningful and important Knowledge Discovery (KD) task.

The well known Least Squares (LS) method is one of the most popular approaches to estimate the parameters in linear regression. But LS has assumptions for which estimators holds some nice and expected properties that do not hold for LR. This shortcoming of LR has lead to the Iterative Re-weighted Least Squares (IRLS) based Maximum Likelihood (ML) method to become popular for estimating the parameters in LR [8, 9]. Pregibon [10] states that "The ML method has good optimality properties in ideal settings, but is extremely sensitive to 'bad' data. 'Bad' from the point of view of outlying responses (Y), and bad from the point of view of extreme points in the design space (X)". Based on the ML approach, LR analysis is sensitive to outliers and gives inaccurate and inconsistent results. One remedy for the outlier problem in regression analysis is the employment of diagnostics. There are many methods used for outlier detection in LR [2, 3, 8, 10]. It is a common idea that outliers are influential to the analysis, but not all the outliers are influential and *vice versa* [11,12]. Hence proper identification and classification of outliers and influential observations are equally important for robust analysis and to make reasonable conclusions. Most of the existing methods are based on single-case deletion and are inefficient in the presence of multiple outliers because of masking and swamping phenomena. We propose an algorithm that can identify multiple influential observations and classify the data into types of outliers in LR.

After fitting an LR model and before making any inference based on the fit, an extra and essential step should be the evaluation and checking of the performance of the model. This paper investigates assessment of the LR results with and without outliers in a dataset, and shows the importance of outlier diagnostics for obtaining reliable

knowledge from modeling, prediction and classification. The paper concentrates on two and three variables data that enables interpretation of the results.

The rest of the paper is arranged as follows: Section II shows how outlier problems are related to reliability issues in KD. In Section III, we give brief ideas about LR, outlier categories, and LR diagnostics. Section IV proposes a new method for the identification of multiple influential observations and a graphical display for the classification of outliers' categories in LR. Section V presents experimental results, evaluates the performance of the proposed technique, shows how outliers can affect the results from LR modeling and classification, and how it relates to reliability issues in KD. In Section VI, we present conclusions.

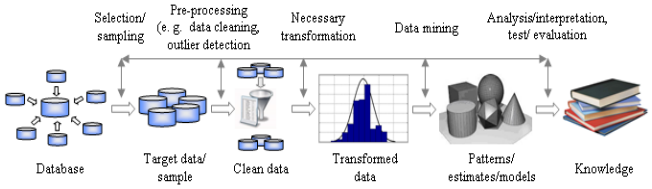

Figure 1. The typical steps constituting the KDD process.

II. Outliers and Reliability Issues

There are many names and definitions of outliers (e.g. abnormal, anomaly, exceptional, intrusion, noise and unusual), and many ways exist for detecting outliers in different fields of knowledge including statistics, machine learning, data mining, computer vision, photogrammetry and remote sensing [7, 10, 13, 14, 15, 16, 17, 18]. Most outlier investigations have been in statistics [2, 3, 13, 19, 20, 21, 22, 23]. The definition of Hawkins [13] captures the meaning and spirit of the word well: "An outlier is an observation that deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism". John [24] defines an outlier as 'surprising veridical data', a point belonging to class A but actually situated inside class B so the true (veridical) classification of the point is surprising to the observer. Outliers occur very frequently in real data, and they often go unnoticed because much data is processed by computers without careful inspection and screening [14]. Outliers may appear because of human error such as keypunch errors, mechanical faults (such as transmission or recoding errors), changes in system behaviour, exceptional events (natural disasters such as earthquakes and floods), instrument error, or simply through natural deviations in populations [14, 17]. The presence of outliers in a dataset may cause the parameter estimation to be erroneous, misclassifying the outcomes and consequently creating problems when making inferences with the wrong model. A key issue which could significantly affect real world applications in data mining is the reliability issues of knowledge discovery [25]. Some interesting and major questions are found in reliability issues: (i) what are the major factors that can make the discovery process unreliable? (ii) how can we make sure that the discovered knowledge are reliable? (iii) under what conditions can a reliable discovery be assured? (iv) what techniques are there that can improve the reliability of discovered knowledge? (v) when can we trust that the discovered knowledge is reliable and reflects the real data? [5, 25, 26, 27]. Answers to these questions are influenced by the presence of outliers in a dataset. We observe in Section V(*B*) that most of the reliability issues can be (directly or indirectly) addressed by the proper investigation and treatment of outliers.

III. Logistic Regression Diagnostics, Classification of Outliers, Related Principles and Methods

A. *Logistic Regression and Diagnostics*

Logistic regression is a technique for describing how a categorical response variable is functionally related with one or more explanatory (predictor/regressor) variables. LR can deal with multinomial as well as binomial response variables. In this paper we focus on the binomial (e.g. 0, 1) response. The customary model for LR is:

$$E(Y \mid X) = \pi(X), \tag{1}$$

where π (X) (hereafter π) is the probability of the outcome (success of an event) of some functionally associated explanatory variable (X). The log of [π (.)/(1–π (.))] can be defined as a linear function called logit (log *odds*) of X, i.e.:

$$g(X) = \ln\left(\frac{\pi}{1-\pi}\right) = \log(odds) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p = X\beta, \tag{2}$$

where
$$\pi = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}, \quad 0 \le \pi \le 1, \tag{3}$$

X is an $n \times k$ ($k = p+1$) data matrix, and $\beta^T=(\beta_0, \beta_1,..., \beta_p)$ is the vector of parameters. Hence the LR model in (1) can be re-written as:

$$Y = \pi + \varepsilon, \tag{4}$$

where Y is a vector of binary (0, 1) responses and ε is the error term:

$$\varepsilon = \begin{cases} 1-\pi & w.p. \quad \pi; \quad if \quad y=1, \\ -\pi & w.p. \quad 1-\pi; \quad if \quad y=0, \end{cases} \tag{5}$$

which follows a distribution with mean 0 and variance $\pi(1–\pi)$. Since ε violates most of the LS assumptions, LR employs the Iterative Re-weighted LS (IRLS) based Maximum Likelihood (ML) method for parameter estimation [8]. It is known that ML estimation is sensitive to outliers. Hence outlier diagnostics are necessary for reliable parameter estimation to fit a correct model and to obtain accurate classification.

Regression diagnostics are quantities computed from the data with the purpose of pinpointing unusual (outliers/influential) observations, which can then be studied and corrected (if necessary) or deleted, followed by fitting the remaining data (inliers) by classical (e.g. LS) methods [14]. The diagnostic approach is a combination of graphical and numerical methods. In recent years diagnostics have become an essential part of LR based study [2, 9, 28].

B. Classification of Outliers in Logistic Regression

As in linear regression, typically outliers in LR can be categorized into three classes: outliers, high leverage points and influential observations. These are obtained by (i) deviation/change in X (explanatory) space, called leverage points (ii) deviation in Y (response variable) not in X, called vertical outliers (iii) deviation in both spaces (Fig. 2). Outliers and high leverage points have a very close relationship to influential observations. Influential observations are defined as points, which either individually or together with several other observations, have a demonstrably larger impact on the calculated values of various estimates (coefficients, standard errors, t-values etc.) [20]. In LR, outliers and influential observations may occur as misclassification between the binary (0, 1) responses. It may occur by meaningful deviation (we also see low leverage) in explanatory variables, which also affect the response as such, so that the usual pattern (S-curve; (Fig. 2 (green line)) of the majority of the data is disrupted [28].

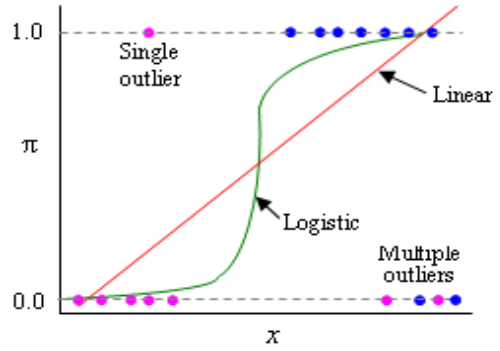

Figure 2. Outliers, and linear and logistic (S-curve) models.

C. Related Diagnostic Principles and Methods

Pregibon [10] provides the foundation of LR diagnostics that has been extended from the idea of linear regression. In LR, the basic building blocks for the identification of outlying and influential points are the residual vector and the projection matrix [10]. In a similar fashion to linear regression, the i^{th} residual $(y_i - \hat{y}_i)$ can be defined in LR as:

$$\hat{\varepsilon}_i = y_i - \hat{\pi}_i . \tag{6}$$

The projection (leverage) matrix is a diagonal matrix that gives the fitted values of the response variable as the projection onto the covariate space. It has been derived by Pregibon [10] as:

$$H = V^{1/2} X (X^T V X)^{-1} X^T V^{1/2} , \tag{7}$$

where V is a diagonal matrix with diagonal elements $v_i = \hat{\pi}_i (1 - \hat{\pi}_i)$. Using Pregibon's linear regression like approximation, (6) holds as $\hat{\varepsilon}_i = y_i - \hat{\pi}_i \approx (1 - h_{ii}) y_i$; h_{ii} is the i^{th} diagonal element of H defined as:

$$h_{ii} = \hat{\pi}_i (1 - \hat{\pi}_i) x_i^T (X^T V X)^{-1} x_i , \tag{8}$$

and variance of the residual $v(\hat{\varepsilon}_i) = v_i (1 - h_{ii})$. Hence, the standardized Pearson residual for LR is defined as:

$$r_{si} = \frac{y_i - \hat{\pi}_i}{\sqrt{v_i (1 - h_{ii})}} . \tag{9}$$

Observations with $| r_{si} | \geq 3$ are generally treated as outliers [8, 9], and large h_{ii} ($> ck/n$; $c = 2$ or 3) values are generally identified as high leverage points. Hosmer and Lemeshow [2] point out that in LR the most extreme points in covariate space may not necessarily have high leverage values if their weights are very small. Residuals, standardized residuals and leverage values are useful for detecting extreme points, but not for assessing their various aspects on the fit [10]. Welsch [29] points out that neither the leverage nor the Studentized residual alone will usually be sufficient to identify the influential cases. The most popular two methods (available in most of the software packages) for the identification of influential observations in linear regression are Cook's Distance (CD) [30] and DFFITS [20]. Variants of them are now available in LR. Welsch [29] suggests DFFITS as a better choice, which is defined in LR as:

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_i^{(-i)}}{v_i^{(-i)} \sqrt{h_{ii}}} . \tag{10}$$

Observations with DFFITS values greater than $3\sqrt{(k/n)}$ are identified as influential cases. All the above measures are based on the single case deletion approach, hence naturally they are affected by the well known masking and swamping phenomena [31] and fail to detect outliers in the presence of multiple outliers and/or influential cases [9, 28]. The group deletion approach introduced in [31] is an approach to outlier detection that forms a clean subset of the data that is presumably free of outliers, and then test the outlyingness of the remaining points relative to the clean subset. A group deleted version of the Generalized Standardized Pearson Residual (GSPR) [9] and the Generalized Weight (GW) [28] have been introduced for detecting multiple outliers and high leverage points respectively in LR. These methods find a suspect group (D) of d outlying/unusual cases with the help of graphical methods, robust techniques such as LMS, LTS, RLS [14] and/or appropriate diagnostics measures like residuals, leverage values and BACON [32]. Interested readers are referred to [9, 28, 31] to find out more about the

identification of the suspect group D and the group deletion approach. The group of remaining (n-d) cases is denoted as R (clean set). Without loss of generality, the data in explanatory variables (X), response variable (Y) and the variance-covariance matrix V can be defined respectively as:

$$X = \begin{bmatrix} X_R \\ X_D \end{bmatrix},\ Y = \begin{bmatrix} Y_R \\ Y_D \end{bmatrix},\ V = \begin{bmatrix} V_R & 0 \\ 0 & V_D \end{bmatrix}. \quad (11)$$

Based on the R set, $\hat{\beta}_{(R)}$ is the corresponding vector of estimated parameters (coefficients). The fitted value for the i^{th} response based on the R set for the entire model can be defined as:

$$\hat{\pi}_{i(R)} = \frac{\exp\left(x_i^T \hat{\beta}_{(R)}\right)}{1+\exp\left(x_i^T \hat{\beta}_{(R)}\right)}. \quad (12)$$

Hence, the i^{th} group-deleted version of residual, corresponding variance and the i^{th} diagonal element of the leverage matrix respectively can be defined as:

$$\hat{\varepsilon}_{i(R)} = y_i - \hat{\pi}_{i(R)},\ v_{i(R)} = \hat{\pi}_{i(R)}(1-\hat{\pi}_{i(R)}), \quad (13)$$

$$h_{ii(R)} = \hat{\pi}_{i(R)}(1-\hat{\pi}_{i(R)}) x_i^T (X_R^T V_R X_R)^{-1} x_i. \quad (14)$$

Using (12, 13, 14), the Generalized Standardized Pearson Residual (GSPR) [9] and the Generalized Weight (GW) [28] are proposed respectively as:

$$r_{si}^* = \begin{cases} \dfrac{y_i - \hat{\pi}_{i(R)}}{\sqrt{v_{i(R)}(1-h_{ii(R)})}} & for\ i \in R, \\ \dfrac{y_i - \hat{\pi}_{i(R)}}{\sqrt{v_{i(R)}(1+h_{ii(R)})}} & for\ i \in D, \end{cases} \quad (15)$$

$$h_{ii}^* = \begin{cases} \dfrac{h_{ii(R)}}{1-h_{ii(R)}} & for\ i \in R, \\ \dfrac{h_{ii(R)}}{1+h_{ii(R)}} & for\ i \in D. \end{cases} \quad (16)$$

Observations with $|r_{si}^*| \geq 3$ and $h_{ii}^* > median(h_{ii}^*) + 3 \times \text{MAD}(h_{ii}^*)$ are identified as outliers and high leverage points respectively.

IV. Proposed Method

This section proposes an influence measure for identifying multiple influential cases in LR. Based on the Mahalanobis Distance (MD), it has the advantage of classifying the observations into outliers, high leverage points, influential and regular observations.

The Mahalanobis Distance (MD) is one of the most well known distances used for identifying multivariate outliers and is defined as:

$$MD_i = \sqrt{(Z_i - \bar{Z})^T \Sigma^{-1} (Z_i - \bar{Z})}, \quad (17)$$

where Z is a m variate random variable with mean $\bar{Z}$ and covariance matrix Σ. It is known that MD follows a Chi-square (χ^2) distribution. Observations with MD values greater than $\sqrt{(\chi^2_{m,0.975})}$ are usually treated as outliers. Due to the non-robustness of $\bar{Z}$ and Σ, MD suffers from masking and swamping phenomena [33]. Although it is still quite easy to detect a single outlier using MD, this approach no longer suffices for multiple outliers [34]. To reduce outlier effects many robust estimators of mean and covariance matrix have been introduced in the literature [34, 35].

We propose a MD type Influence Distance (ID) to find influential cases in LR. We generate a two column matrix of GSPR in (15) and GW in (16) to preserve information about both generalized standardized Pearson residuals and generalized leverage values. We define this matrix as the generalized residual-leverage matrix G:

$$G = [r_{si}^* \quad h_{ii}^*]. \quad (18)$$

We define ID as:

$$ID_i = \sqrt{\left(G_i - \bar{G}_{(R)}\right)^T \Sigma_{(R)}^{-1} \left(G_i - \bar{G}_{(R)}\right)}, \quad (19)$$

where $\bar{G}_{(R)}$ and $\Sigma_{(R)}$ are the mean and covariance matrix based on the R group (excluding the observations which are identified as outliers by GSPR). To reduce the effects of outliers on $\bar{Z}$ and Σ in (17), and to make the ID robust, we use the $\bar{G}_{(R)}$ and $\Sigma_{(R)}$ in (19). In a similar fashion to MD, ID follows a Chi-square distribution, and observations having ID values larger than $\sqrt{(\chi^2_{2,0.975})} = 2.716$ are identified as influential observations.

We also propose a classification plot using the ID to classify the data into outliers, high leverage points, influential and regular observations, see Figure 3. We generate a scatter plot of GSPR (r_{si}^*) versus GW (h_{ii}^*) and sketch three cut-off lines. The first two cut-off lines are horizontal, parallel to the h_{ii}^* axis at GSPR= ±3 to find outliers, and the third cut-off line is vertical, parallel to the r_{si}^* axis at $median(h_{ii}^*) + 3 \times \text{MAD}(h_{ii}^*)$ that can separate the high leverage points from the usual cases. Finally, we draw a confidence (influence) ellipse on the scatter plot based on the ID values in (19). Finally, observations outside the ellipse are identified as influential cases. The remaining observations that are not outliers, high leverage points or influential observations can be treated as regular or usual cases. The identification and classification methods are summarized in Algorithm 1.

Algorithm 1.

(i) Calculate r_{si}^* in (15) and h_{ii}^* in (16) using the group deletion approach in Section III (*C*).
(ii) Construct the matrix *G* according to (18).
(iii) Calculate $\overline{G}_{(R)}$ and $\Sigma_{(R)}$ based on the *R* group after the deletion of outlying cases identified by r_{si}^*.
(iv) Calculate ID using (19).
(v) Find influential observations for which

$$ID_i > \chi^2_{2,0.975} = 2.716 .$$

(vi) To sketch the classification plot (Fig. 3): (a) draw an scatter plot r_{si}^* versus h_{ii}^* (b) draw cut-off lines at ±3 and *median*(h_{ii}^*)+3×MAD(h_{ii}^*) through the r_{si}^* and h_{ii}^* axes respectively (c) draw an influence ellipse based on the ID values and the Chi-square cut-off value.

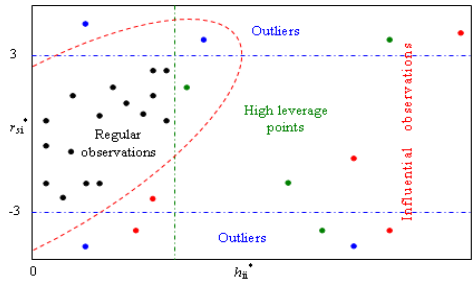

Figure 3. Classification plot.

V. Experimental Results

We consider numerical examples to assess the efficiency of the proposed algorithm for the identification of influential observations through well-referred real datasets. Second, we investigate the problems of outliers in model building and classification, and the necessities of outlier diagnostics for inferring reliable knowledge from LR as a part of the KD.

A. Performance Evaluation of the Algorithm

To determine the efficiency of the proposed algorithm for finding multiple influential cases in LR, we use the two following datasets in our demonstration.

1) Modified Brown Data:

We first consider the Brown et al. [36] data. The original dataset contains 53 observations of five regressors. We only consider acid phosphatase (A.P.) as a regressor here for illustrating diagnostics in simple LR. The original objective was to see whether an elevated level of phosphatase in blood serum would be of value for predicting whether or not a prostate cancer patient also had lymph node involvement (L.N.I.) [8]. Ryan [8] shows that the original data contains a single (case 24) outlier. To see the influence of the presence of multiple outliers, and to see the effects of masking and swamping, we modify the data by including two more cases: 54 and 55, and changing the responses (1 to 0) of the 25th and 53rd patients. The modified dataset on the 55 patients are given in Table 1. Although the X-Y scatter plot in LR has a very little explorer value, the scatter plot (Fig. 4(a)) of the L.N.I. versus A.P. values shows that the five cases (24, 25, 53, 54, 55) are among the patients without nodal involvement that have much higher A.P. levels than the majority of cases.

TABLE. 1 MODIFIED BROWN DATA

i	L.N.I.	A.P.	*i*	L.N.I.	A.P.	*i*	L.N.I.	A.P.	*i*	L.N.I.	A.P.
1	0	48	15	0	47	29	0	50	43	1	81
2	0	56	16	0	49	30	0	40	44	1	76
3	0	50	17	0	50	31	0	55	45	1	70
4	0	52	18	0	78	32	0	59	46	1	78
5	0	50	19	0	83	33	1	48	47	1	70
6	0	49	20	0	98	34	1	51	48	1	67
7	0	46	21	0	52	35	1	49	49	1	82
8	0	62	22	0	75	36	0	48	50	1	67
9	1	56	23	1	99	37	0	63	51	1	72
10	0	55	**24**	**0**	**187**	38	0	102	52	1	89
11	0	62	**25**	**(1)0**	**136**	39	0	76	**53**	**(1)0**	**126**
12	0	71	26	1	82	40	0	95	**54**	**0**	**200**
13	0	65	27	0	40	41	0	66	**55**	**0**	**220**
14	1	67	28	0	50	42	1	84			

Table 2 contains all the single and group deletion diagnostics results. They show that standardized Pearson residuals fail to find any of the outliers. This means all the five outliers are masked as good points. Leverage values identify four cases (24, 25, 54, 55) as high leverage points. To employ the proposed algorithm for finding multiple influential cases, we form the deletion group *D* of the five suspect cases determined by Fig. 4(a). Based on the clean set *R*, we calculate GSPR [9] and GW [28]. Results of GSPR (r_{si}^*) in Table 2 show that three cases (24, 54, 55) are correctly identified as outliers (Fig. 4(e)) and six cases (20, 23, 25, 38, 40, 53) are identified as high leverage points by the group deleted version of leverage values, GW (h_{ii}^*, Fig. 4(f)). Single case deletion influence measure DFFITS failed to detect any of the cases as influential, which shows that DFFITS values are affected by masking. The results in table 2 (columns 7 and 14) along with Fig. 4(g) shows that the proposed ID successfully identifies all the five influential cases (24, 25, 53, 54, 55) as well as one more (case 38) which has a large A.P. value (102). It is clear that case 38 was masked before by the presence of multiple unusual cases. Fig. 4(h) shows the overall classification properly.

2) Modified Finney Data

We consider another dataset from Finney [37], which has been extensively analyzed later by many authors as an example of multiple LR diagnostics [9, 10, 28]. This dataset has 39 cases with two regressors. The dataset in Table 3 was obtained in a controlled study of the effect of rate and volume of air inspired on a transient vasoconstriction in the skin of the digits [10]. The character plot of Fig. 5(a) has been created by plotting rate versus volume, and the characteristics corresponding to occurrence (1) and non occurrence (0) have been shown by different colors. Using a contour plot Pregibon [10] shows that this dataset may contain two outliers (cases 4 and 18). To get more outliers and to see the effects of multiple outliers, we deliberately interchange (0 to 1) for cases (10, 11) as shown in Table 3.

TABLE. 2 DIAGNOSTIC RESULTS FOR MODIFIED BROWN DATA

i	$\|r_{si}\|$ (3.00)	h_{ii} (0.073)	$\|DFFITS_i\|$ (0.572)	$\|r_{si}^*\|$ (3.00)	h_{ii}^* (0.081)	ID_i (2.716)	*i*	$\|r_{si}\|$ (3.00)	h_{ii} (0.073)	$\|DFFITS_i\|$ (0.572)	$\|r_{si}^*\|$ (3.00)	h_{ii}^* (0.081)	ID_i (2.716)
1	-0.740	0.029	-0.124	-0.520	0.040	0.578	29	-0.737	0.027	-0.120	-0.543	0.037	0.655
2	-0.728	0.023	-0.110	-0.615	0.029	0.906	30	-0.753	0.036	-0.141	-0.441	0.052	0.459
3	-0.737	0.027	-0.120	-0.543	0.037	0.655	31	-0.729	0.024	-0.111	-0.602	0.030	0.866
4	-0.734	0.026	-0.116	-0.566	0.034	0.739	32	-0.723	0.022	-0.106	-0.655	0.026	1.017
5	-0.737	0.027	-0.120	-0.543	0.037	0.655	33	1.391	0.029	0.237	1.998	0.040	1.994
6	-0.739	0.028	-0.122	-0.531	0.038	0.615	34	1.396	0.026	0.228	1.869	0.035	1.863
7	-0.743	0.030	-0.128	-0.499	0.043	0.515	35	1.393	0.028	0.234	1.954	0.038	1.948
8	-0.719	0.020	-0.102	-0.698	0.024	1.104	36	-0.740	0.029	-0.124	-0.520	0.040	0.578
9	1.406	0.023	0.216	1.672	0.029	1.696	37	-0.718	0.020	-0.101	-0.714	0.023	1.126
10	-0.729	0.024	-0.111	-0.602	0.030	0.866	38	-0.674	0.030	-0.111	-1.740	**0.131**	**3.173**
11	-0.719	0.020	-0.102	-0.698	0.024	1.104	39	-0.701	0.018	-0.093	-0.949	0.034	1.091
12	-0.707	0.018	-0.095	-0.849	0.026	1.165	40	-0.681	0.025	-0.103	-1.476	**0.101**	2.215
13	-0.715	0.019	-0.099	-0.745	0.023	1.159	41	-0.714	0.019	-0.098	-0.761	0.023	1.170
14	1.431	0.019	0.198	1.315	0.023	1.428	42	1.475	0.020	0.204	0.926	0.056	1.054
15	-0.742	0.029	-0.126	-0.510	0.042	0.545	43	1.466	0.019	0.200	0.983	0.046	0.999
16	-0.739	0.028	-0.122	-0.531	0.038	0.615	44	1.453	0.018	0.196	1.089	0.034	1.113
17	-0.737	0.027	-0.120	-0.543	0.037	0.655	45	1.438	0.018	0.196	1.234	0.025	1.335
18	-0.699	0.019	-0.093	-0.993	0.038	1.063	46	1.458	0.019	0.197	1.045	0.038	1.048
19	-0.693	0.020	-0.094	-1.114	0.052	1.101	47	1.438	0.018	0.196	1.234	0.025	1.335
20	-0.678	0.027	-0.106	-1.584	**0.114**	2.624	48	1.431	0.019	0.198	1.315	0.023	1.428
21	-0.734	0.026	-0.116	-0.566	0.034	0.739	49	1.469	0.019	0.201	0.964	0.049	1.003
22	-0.702	0.018	-0.094	-0.928	0.032	1.108	50	1.431	0.019	0.198	1.315	0.023	1.428
23	1.520	0.028	0.244	0.689	**0.118**	2.706	51	1.443	0.018	0.195	1.183	0.027	1.264
24	-0.634	**0.186**	-0.275	**-9.979**	0.051	**9.995**	52	1.489	0.022	0.214	0.839	0.075	1.420
25	-0.650	**0.074**	-0.167	0.274	**0.126**	**2.837**	**53**	-0.656	0.058	-0.148	0.339	**0.132**	**3.061**
26	1.469	0.019	0.201	0.964	0.049	1.003	**54**	-0.634	**0.222**	-0.308	**-13.311**	0.037	**13.426**
27	-0.753	0.036	-0.141	-0.441	0.052	0.459	**55**	-0.637	**0.281**	-0.366	**-20.662**	0.021	**20.914**
28	-0.737	0.027	-0.120	-0.543	0.037	0.655							

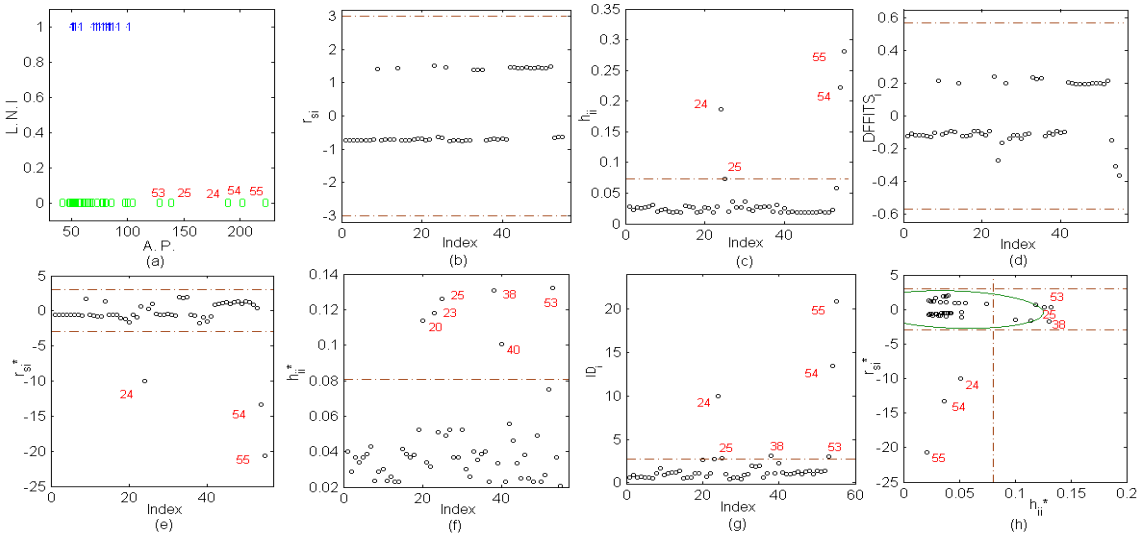

Figure 4. Modified Brown data (a) scatter plot; L.N.I. versus A.P. (b) index plot of standardized Pearson residual (c) index plot of leverage values (d) index plot of DFFITS (e) index plot of GSPR (f) index plot of GW (g) index plot of ID (h) classification plot.

Diagnostics results in Table 4 show that Pearson standardized residuals fail to identify any outlier, all the four outliers are masked in the presence of the group of outliers (cases 4, 8, 10 and 11). Leverage values identify two wrong cases (31, 32) as high leverage points, while DFFITS fails to identify any one of the influential cases. We calculate the group deletion diagnostics (GSPR and GW) measures using the same suspect group (cases 4, 10, 11 and 18) suggested by Imon and Hadi [9]. In Table 4, we see the GSPR values (columns 5 and 12; Fig. 5(e)) identifies the four cases as outliers and the GW values (columns 6 and 13) identify many cases as high leverage points (Fig. 5(f)).

TABLE. 3 MODIFIED FINNEY DATA

i	Y	Volume	Rate	*i*	Y	Volume	Rate	*i*	Y	Volume	Rate
1	1	3.70	0.825	14	1	1.40	2.330	27	1	1.80	1.500
2	1	3.50	1.090	15	1	0.75	3.750	28	0	0.95	1.900
3	1	1.25	2.500	16	1	2.30	1.640	29	1	1.90	0.950
4	**1**	**0.75**	**1.500**	17	1	3.20	1.600	30	0	1.60	0.400
5	1	0.80	3.200	**18**	**1**	**0.85**	**1.415**	31	1	2.70	0.750
6	1	0.70	3.500	19	0	1.70	1.060	32	0	2.35	0.030
7	0	0.60	0.750	20	1	1.80	1.800	33	0	1.10	1.830
8	0	1.10	1.700	21	0	0.40	2.000	34	1	1.10	2.200
9	0	0.90	0.750	22	0	0.95	1.360	35	1	1.20	2.000
10	**(0)1**	**0.90**	**0.450**	23	0	1.35	1.350	36	1	0.80	3.330
11	**(0)1**	**0.80**	**0.570**	24	0	1.50	1.360	37	0	0.95	1.900
12	0	0.55	2.750	25	1	1.60	1.780	38	0	0.75	1.900
13	0	0.60	3.000	26	0	0.60	1.500	39	1	1.30	1.625

To sort out the influential cases, we employ the algorithm for ID. We use the suspect group (cases 4, 10, 11 and 18) identified by GSPR for the deletion, and compute the mean and covariance matrix from the remaining 35 cases, and the ID values using (19). Results in Table 4 (columns 7 and 14) show that ID identifies all four cases (4, 10, 11 and 18) with three more cases (13, 32, 39) as influential that were masked before. Since the ID values for the influential cases are much larger than the values of the non-influential cases, we rescale the ID values by taking logs for better visualization. Fig. 5(g) shows clear separation between influential and non-influential cases. The character plot (Fig. 5(a)) of the cases (13, 32, 39) justifies the performance of the proposed ID.

TABLE. 4 DIAGNOSTIC RESULTS OF MODIFIED FINNEY DATA

i	$\lvert r_{si}\rvert$ (3.00)	h_{ii} (0.154)	$\lvert DFFITS_i\rvert$ (0.832)	$\lvert r_{si}^*\rvert$ (3.00)	h_{ii}^* (0.044)	ID_i (2.716)	*i*	$\lvert r_{si}\rvert$ (3.00)	h_{ii} (0.154)	$\lvert DFFITS_i\rvert$ (0.832)	$\lvert r_{si}^*\rvert$ (3.00)	h_{ii}^* (0.044)	ID_i (2.716)
1	0.1491	0.072	-0.127	0.000	0.00000	0.532	21	-0.6157	0.100	-0.201	-0.001	0.00007	0.533
2	0.1543	0.068	-0.112	0.000	0.00000	0.532	22	-0.7032	0.061	-0.175	-0.005	0.00068	0.532
3	0.5799	0.061	0.148	0.035	0.01708	0.444	23	-1.0176	0.045	-0.194	-0.151	**0.09264**	0.331
4	1.6914	0.071	0.353	**587.164**	0.00013	**965.085**	24	-1.1836	0.048	-0.210	-0.615	**0.17485**	0.937
5	0.6074	0.118	0.209	0.041	0.02304	0.417	25	0.6301	0.052	0.134	0.079	**0.05443**	0.288
6	0.5683	0.149	0.215	0.020	0.00763	0.489	26	-0.5520	0.084	-0.166	0.000	0.00001	0.533
7	-0.3570	0.093	-0.099	0.000	0.00000	0.532	27	0.6174	0.067	0.143	0.061	0.03996	0.344
8	-0.9807	0.039	-0.190	-0.110	**0.06880**	0.357	28	-0.9577	0.044	-0.201	-0.086	**0.05323**	0.389
9	-0.4759	0.094	-0.150	0.000	0.00000	0.532	29	0.7854	0.096	0.222	0.576	**0.42628**	1.829
10	2.7516	0.102	0.790	**44522.925**	0.00000	**73194.797**	30	-0.7693	0.128	-0.244	-0.008	0.00156	0.531
11	2.8168	0.098	0.774	**56039.735**	0.00000	**92128.254**	31	0.4223	**0.160**	0.184	0.001	0.00009	0.531
12	-1.1087	0.110	-0.359	-0.276	**0.27368**	0.572	32	-1.3568	**0.288**	-0.444	-1.144	**1.28008**	**4.314**
13	-1.3554	0.124	-0.459	-1.997	**0.90055**	**3.504**	33	-1.0578	0.037	-0.200	-0.229	**0.14620**	0.342
14	0.5540	0.058	0.134	0.023	0.00941	0.480	34	0.7921	0.045	0.171	0.731	**0.30392**	1.639
15	0.4682	0.151	0.171	0.003	0.00043	0.528	35	0.8071	0.039	0.161	0.870	**0.24012**	1.678
16	0.3559	0.088	0.096	0.000	0.00001	0.532	36	0.5649	0.126	0.201	0.020	0.00801	0.488
17	0.1516	0.057	-0.127	0.000	0.00000	0.532	37	-0.9577	0.044	-0.201	-0.086	**0.05323**	0.389
18	1.6115	0.066	0.335	**386.514**	0.00026	**635.219**	38	-0.7972	0.061	-0.198	-0.015	0.00438	0.524
19	-1.2160	0.073	-0.243	-0.718	**0.21561**	1.072	39	0.9132	0.037	0.173	2.658	**0.18395**	**4.477**
20	0.5176	0.066	0.122	0.012	0.00351	0.511							

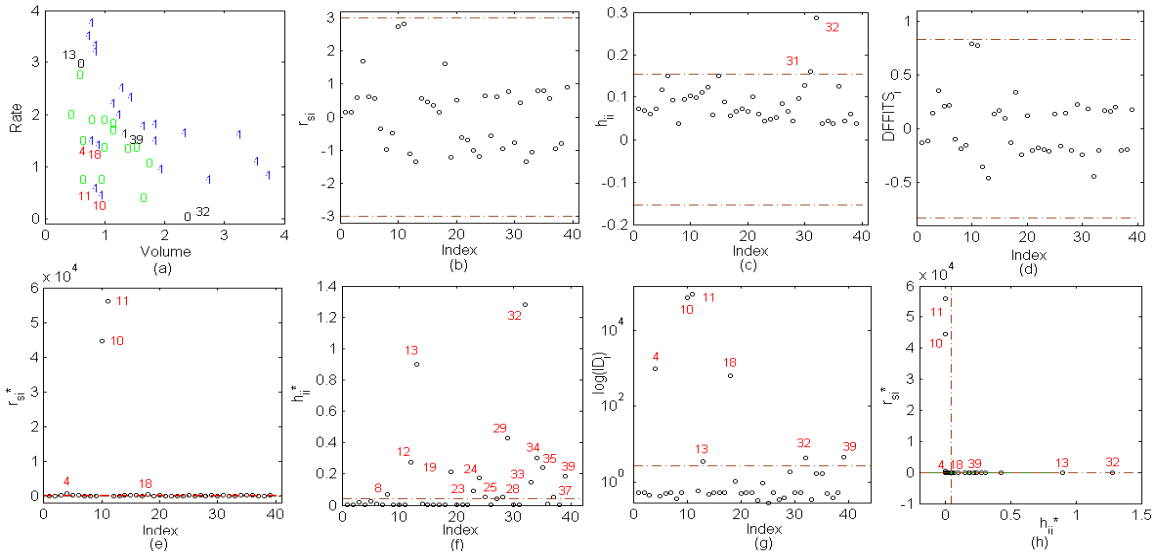

Figure 5. Modified Finney data (a) character plot; rate versus volume with the response values (1, 0) (b) index plot of standardized Pearson residual (c) index plot of leverage values (d) index plot of DFFITS (e) index plot of GSPR (f) index plot of GW (g) index plot of ID (h) classification plot.

The huge differences among the values of the diagnostics measures mean the confidence ellipse is not clearly visible, the graphical plot performs the classification well.

In summary, results from the above two datasets show that although GSPR [9] and GW [28] are able to identify outliers and high leverage points respectively, they can not identify influential cases, whereas the proposed ID perfectly identifies the influential cases even in the presence of multiple unusual cases and free from masking and swamping effects. Moreover, ID has the ability to classify the outliers' categories and regular observations as well.

B. LR Model Reliability Checking and Performance Evaluation in Prediction and Classification

This section analyzes the LR model for the Brown [36] dataset. The analysis is carried out for the datasets with and without outliers detected by the proposed algorithm. Table 5 contains results for LR models from the dataset with and without outliers to allow comparison and to see the effects created in the presence of outliers. Results are analyzed for statistical significance both for estimation of the parameters and testing of the estimates that can be directly related to reliability issues in KD. We perform the necessary tasks as: (i) parameter estimation and model fitting, (ii) statistical significance of the individual parameters estimated, (iii) overall model evaluation, (iv) goodness-of-fit test, (v) classification power evaluation of the LR classifier, and (vi) power and accuracy evaluation of the prediction.

For the results in Table 5, we fit the logit model to the data with and without outliers respectively as:

$$\log(\pi/(1-\pi)) = -0.463 - 0.003 \times A.P., \quad (20)$$

$$\log(\pi/(1-\pi)) = -4.134 + 0.055 \times A.P. \quad (21)$$

Using (20, 21), we see big shifts among the parameter values of the fitted models. The extreme case is the gradient is reversed. According to the Z-test both the estimated parameters (slope and intercept) are significant at the 5% level (i.e. $p < 0.05$) for the LR model from the data without outliers. In the presence of outliers, the same parameters are insignificant to the fitted model. A similar decision can be made for the hypothesis: all slopes are zero. That means that in the presence of outliers, the LR model is unreliable for prediction ($p = 0.669$). Without outliers the LR model is highly significant with $p = 0.007$. Fig 6 shows the predicted probabilities from the derived models. The LR model from observations with outliers does not follow any S-curve (red line) shape but after the deletion of the outliers the predicted probabilities produce a well shaped S-curve (green).

To test the overall model, we use both inferential and descriptive statistics. For the Hosmer-Lemshow (H-L) goodness-of-fit test, the p value is larger than 0.05 which is good for a well fitted model. This desirable non-significance outcome indicates that the model prediction significantly supports the observed. The same statistic draws the reverse conclusion when the dataset is outlier contaminated. We know that a better model is one that results in a larger likelihood, where the likelihood is the joint probability of observing the sample values given the model parameters. This measures the fitted model's success rate for explaining the response variable by the predictor(s). Log likelihood (LL) and -2 Log likelihood (-2LL) values show that the LR model without outliers fits better than the model with outliers. Similar outcomes can be drawn from the variants of the coefficient of determination (R^2). The R^2 values of Cox and Snell [38] and Nagelkerke [39] are much larger for the model without outliers. Results of R^2 show that after deleting the outliers the explanation ability of the LR model has been significantly increased.

LR can classify responses according to the respective predicted probabilities using a contingency (classification) table. In Table 6, the 2×2 contingency table shows that the LR model for all observations makes 58% overall correct classification which is increased to 69% after the deletion of outliers. Results also show that the LR model totally fails to classify the response (L.N.I. involvement;1) before the deletion of outliers. Respective mosaic plots (Fig. 7) show the power of classification for both types of responses.

TABLE 5. LR MODEL FIT AND SIGNIFICANCE TEST

Results for all observations								**Results without outliers**						
Parameter estimation								***Parameter estimation***						
Predictor	Coef.	S. E.	Z	*P*	Odds Ratio	95% Conf. Int.		Coef.	S. E.	Z	*P*	Odds Ratio	95% Conf. Int.	
						Lower	Upper						Lower	Upper
Constant	-0.463	0.674	-0.69	0.492				-4.134	1.486	-2.78	0.005			
A.P.	-0.003	0.008	-0.42	0.677	1.00	0.98	1.01	0.055	0.022	2.51	0.012	1.06	1.01	1.10
Test								***Test***						
Test that all slopes are zero: G = 0.183, df = 1, *P*-Value = 0.669								Test that all slopes are zero: G = 7.31, df = 1, *P*-Value = 0.007						
Goodness -of-fit test								***Goodness -of-fit test***						
				χ^2	df	*p*					χ^2	df	*p*	
Pearson				42.144	34	0.159		Pearson			33.295	28	0.225	
Deviance				53.407	34	0.018		Deviance			41.167	28	0.052	
Hosmer-Lemeshow (H-L)				23.059	8	0.003		Hosmer-Lemeshow (H-L)			6.044	8	0.642	
Model Summery								***Model Summery***						
Log-Likelihood (LL)		-2 LL		Cox & Snell R^2		Nagelkerke R^2		Log-Likelihood (LL)		-2 LL	Cox & Snell R^2		Nagelkerke R^2	
-34.681		69.363		0.003		0.005		-28.562		57.123	0.139		0.190	

Assessment of the predictive ability is vital and a critical step for the tools and algorithms that can predict. We use the well known Receiver Operating Characteristic (ROC) curve to assess the predictive ability of the LR models. This curve generates a number of classifications with different cut-off values between 0 and 1, and calculates the sensitivity (true positive) and specificity (true negative) for each of the cut-off values. Plotting the sensitivity against (1 – specificity) creates the desired curve and the Area Under the Curve (AUC), is used to assess the predictive ability of the underlying method. We see AUCs of the ROC curves generated by the predicted LR models from the datasets with and without outliers in Fig. 8(a) and Fig. 8(b) respectively. Table 7 shows the AUC for the LR model in the presence of outliers is 0.22 (insignificant predictive ability of the model) which is increased to 0.78 (highly significant) when the outliers have been deleted.

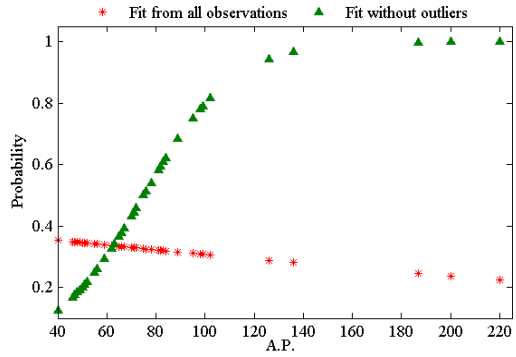

Figure 6. Predicted probabilities versus A.P.

TABLE 6. CLASSIFICATION RESULTS WITH OUTLIERS AND WITHOUT OUTLIERS

Predicted status	**All observations**			**Without outliers**		
	Actual Status			***Actual status***		
	Absence (0)	Presence (1)	Total	Absence (0)	Presence (1)	Total
Absence (0)	32 (58.18%)	23 (41.82 %)	55 (100%)	25 (45.45%)	10 (18.18%)	35 (63.64%)
Presence (1)	0 (0%)	0 (0%)	0 (0%)	7 (12.73%)	13 (23.64%)	20 (36.36%)
Total	32 (58.18%)	23 (41.82 %)	55	32 (58.18%)	23 (41.82%)	55
Correct classif.	58.18%			69.09%		

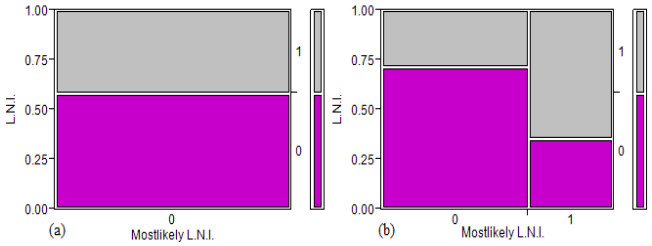

Figure 7. Mosaic plot (a) classification with outliers (b) classification without outliers.

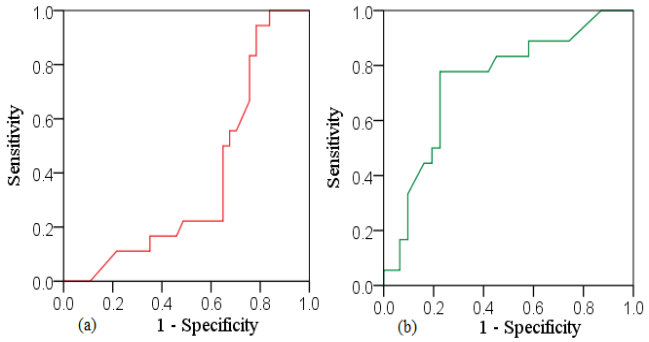

Figure 8. ROC curve (a) data with outliers (b) data without outliers.

TABLE 7. ROC CURVES RESULTS

	Area	**S. E.**	**Sig. (*p*)**	**Asymptotic 95% confidence interval**	
				Lower Bound	Upper Bound
All observations	0.219	0.064	0.000	0.094	0.345
Without outliers	0.781	0.064	0.000	0.655	0.906

VI. CONCLUSIONS

This paper proposes a diagnostic measure for identifying multiple influential observations in logistic regression. It introduces a classification graph to classify outliers, high leverage points and influential observations in the same plot at one time. Diagnostic results show that the proposed measure efficiently identifies multiple influential cases, and the graph is helpful for visualizing outlier categories. In this paper, we discussed the results from logistic regression models with and without outliers. Results show that without careful outlier investigation, it may not be possible to get reliable knowledge using logistic regression for predictive modeling and classification. We observe that the outlier investigation in logistic regression is highly related to the issues raised for reliable knowledge discovery in Section II. We can answer the issues: (i) outlier detection is one of the major factors that affect the reliability of the discovery process, (ii) the conditions for reliable knowledge discovery can be improved by parameter estimation and testing the significance of the estimates, (iii) proper outlier diagnostics and treatment (deletion or correction of the outlying observations) can improve the reliability of discovered knowledge, (iv) when the test results meet the required statistical significance level, then we can trust that the discovered knowledge is reliable and reflects the real data. Therefore outlier detection and proper treatment is vital for obtaining reliable knowledge, and should be considered as a data preprocessing step in knowledge discovery in databases (KDD). The proposed diagnostic method is introduced for the binomial response variable in logistic regression. Future research will investigate the diagnostic method for (i) multinomial response variables, and (ii) large and high dimensional data as higher dimensional data presents extra problems that need to be addressed.

ACKNOWLEDGMENT

This study has been carried out as a PhD research supported by a Curtin University International Postgraduate Research Scholarship (IPRS). The work has also been supported by the Cooperative Research Centre for Spatial Information, whose activities are funded by the Australian Commonwealth's Cooperative Research Centres Programme.

REFERENCES

[1] J. F. Elder, and D. Pregibon, "A Statistical Perspective on KDD," Proc. of the KDD-95, 1995, pp. 87 – 93.

[2] D. W. Hosmer, and S. Lemeshow, Applied Logistic Regression, 2nd Ed., New York: John Wiley and Sons, 2000.

[3] S. Chatterjee, and A. S. Hadi, Regression Analysis by Examples, 4th Ed., New York: John Wiley and Sons, 2006.

[4] P. Komarek and A. W. Moore, "Making Logistic Regression A Core Data Mining Tool," Robotics Institute, Paper 218, 2005, http://repository.cmu.edu/robotics/218

[5] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, "The KDD Process for Extracting Useful Knowledge from Volumes of Data," Communications of the ACM, Vol. 39 (11), 1996, pp. 27 – 34.

[6] K. Collier, B. Carey, E. Grusy, C. Marajaniemi, and D. Sautter, A Perspective on Data Mining, 1998, http://insight.nau.edu/downloads/dm%20perspective%20v2.pdf, Accessed: 22– 07– 2012.

[7] E. M. Knorr, R. T. Ng, and V. Tucakov, "Distance-based Outlier: Algorithms and Applications," The International Journal of Very Large Databases, Vol. 8(3-4), 2000, pp. 237 – 253.

[8] T. P. Ryan, Modern Regression Methods, New York: John Wiley and Sons, 1997.

[9] A. H. M. R. Imon, and A. S. Hadi, "Identification of Multiple Outliers in Logistic Regression," Communications in Statistics-Theory and Methods, Vol. 37, 2008, pp. 1 – 13.

[10] D. Pregibon, "Logistic Regression Diagnostics," Annals of Statistics, Vol. 9, 1981, pp. 977 – 986.

[11] D. F. Andrews, and D. Pregibon, "Finding the Outliers that Matter," Journal of the Royal Statistical Society, Series-B, Vol. 40, 1978, pp. 85 – 93.

[12] S. Chatterjee, and A. S. Hadi, "Influential Observations, High Leverage Points, and Outliers in Regression," Statistical Sciences, Vol. 1(3), 1986, pp. 379 – 416.

[13] D. M. Hawkins, Identification of Outliers, London: Chapman and Hall, 1980.

[14] P. J. Rousseeuw, and A. M. Leroy, Robust Regression and Outlier Detection. New York: John Wiley and Sons, 1987.

[15] M. Breunig, H. P. Kriegel, R. Ng, and J. L. O. F. Sander, "Identifying Density-based Local Outliers," Proc. of the ACM SIGMOD, International Conference on Management of Data, New York: ACM Press, 2000, pp. 93 – 104.

[16] C. C. Aggarwal, and P. S. Yu, "Outlier Detection for High Dimensional Data," Proc. of the 2001 ACM SIGMOD International Conference on Management of Data, ACM SIGMOD Record, Vol. 30(2), 2001, pp. 37 – 46.

[17] V. J. Hodges, and J. Austin, "A Survey of Outlier Detection Methodologies," Artificial Intelligence Review, Vol. 22, 2004, pp. 85 – 126.

[18] S. Sotoodeh, "Outlier Detection in Laser Scanner Point Clouds," Proc. of the IAPRS, Dresden, Vol. XXXVI/5, 2006, pp. 297 – 301.

[19] V. Barnett, and T. B. Lewis, Outliers in Statistical Data, 3rd Ed., New York: John Wiley and Sons, 1995.

[20] D. A. Belsley, E. Kuh, and R. E. Welsch, Regression Diagnostics: Identifying Influential Data and Sources of Colinearity, New York: John Wiley and Sons, 1980.

[21] R. D. Cook, and S. Weisberg, Residuals and Influence in Regression, London: Chapman and Hall, 1982.

[22] P. J. Rousseeuw and M. Hubert, "Robust Statistics for Outlier Detection," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol. 1(1), 2011, pp. 73 – 79.

[23] A. A. M. Nurunnabi, A. H. M. R. Imon and M. Nasser, "A Diagnostic Measure for Influential Observations in Linear Regression," Communications in Statistics-Theory and Methods, Vol. 40(7), 2011, pp. 1169 – 1183.

[24] G. H. John, "Robust Decision Trees: Removing Outliers from Databases," Proc. of The First International Conference on Knowledge Discovery and Data Mining, Menlo Park, CA : AAAI Press, 1995, pp. 174 – 179.

[25] H. Dai, J. N. K. Liu, and E. Smirnov, Reliable Knowledge Discovery, New York: Springer, 2012.

[26] Y. Feng, and Z. Wu, "Enhancing Reliability Throughout Knowledge Discovery Process," Proc. of the 1st International Workshop on Reliability Issues in Knowledge Discovery, Hong Kong, China, 2006.

[27] A. A. M. Nurunnabi, and H. Dai, Robust-Diagnostic Regression: A Prelude for Inducing Reliable Knowledge from Regression, In Reliable Knowledge Discovery, Eds. H. Dai, J. N. K. Liu, and E. Smirnov (Eds.), New York: Springer, 2012, pp. 69 – 90.

[28] A. A. M. Nurunnabi, A. H. M. R. Imon, and M. Nasser, "Identification of Multiple Influential Observations in Logistic Regression," Journal of Applied Statistics, Vol. 37(10), 2010, pp. 1605 – 1624.

[29] R. E. Welsch, "Influence Functions and Regression Diagnostics," In: R. L. Launar, and A. F. Siegel, (Eds.), Modern Data Analysis, New York: Academic Press, 1982.

[30] R. D. Cook, "Detection of Influential Observations in Linear Regression," Technometrics, Vol. 19, 1977, pp. 15 – 18.

[31] A. S. Hadi, and J. S. Simonoff, "Procedures for the Identification of Outliers," Journal of the American Statistical Association, Vol. 88, 1993, pp. 1264 – 1272.

[32] N. Billor, A. S. Hadi and F. Velleman, "BACON: Blocked Adaptive Computationally Efficient Outlier Nominator," Computational Statistics and Data Analysis, Vol. 34, 2000, pp. 279 – 298.

[33] P. J. Rousseeuw and V. Zomeren, "Unmasking Multivariate Outliers and Leverage Points," Journal of the American Statistical Association, Vol. 85 (411), 1990, pp. 633 – 639.

[34] P. J. Rousseeuw and van Driessen, "A Fast Algorithm for the Minimum Covariance Determinant Estimator," Technometrics, Vol. 41(3), 1999, pp. 212 – 223.

[35] R. A. Maronna and V. J. Yohai, "Robust Estimation of Multivariate Location and Scatter," Encyclopedia of Statistics, Vol. 2, New York: John Wiley and Sons, 1998.

[36] B. W. Jr. Brown, "Prediction Analysis for Binary Data," In: R. J. Jr. Miller, B. Efron, B. W. Jr. Brown and L. E. (Eds.), Biostatistics Casebook, New York: John Wiley and Sons, 1980.

[37] D. J. Finney, "The Estimation From Individual Records of the Relationship between Dose and Quantal Response," Biometrika, Vol. 34, 1947, pp. 320 – 334.

[38] D. R. Cox and E. J. Snell, "The Analysis of Binary Data," 2nd Ed., London: Chapman and Hall, 1989.

[39] N. J. D. Nagelkerke, "A Note on a General Definition of the Coefficient of Determination," Biometrika, Vol. 78, 1991, pp. 691– 692.