Imbalance Effects on Classification Using Binary Logistic Regression

Hezlin Aryani Abd Rahman[1(✉)] and Bee Wah Yap[2]

[1] Faculty of Computer and Mathematical Sciences, Centre of Statistical and Decision Science Studies, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
hezlin@tmsk.uitm.edu.my

[2] Faculty of Computer and Mathematical Sciences, Advanced Analytics Engineering Centre, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
beewah@tmsk.uitm.edu.my

Abstract. Classification problems involving imbalance data will affect the performance of classifiers. In predictive analytics, logistic regression is a statistical technique which is often used as a benchmark when other classifiers, such as Naïve Bayes, decision tree, artificial neural network and support vector machine, are applied to a classification problem. This study investigates the effect of imbalanced ratio in the response variable on the parameter estimate of the binary logistic regression via a simulation study. Datasets were simulated with controlled different percentages of imbalance ratio (IR), from 1 % to 50 %, and for various sample sizes. The simulated datasets were then modeled using binary logistic regression. The bias in the estimates was measured using MSE (Mean Square Error). The simulation results provided evidence that imbalance ratio affects the parameter estimates where severe imbalance (IR = 1 %, 2 %, 5 %) has higher MSE. Additionally, the effects of high imbalance (IR $\leq$ 5 %) will be more severe when sample size is small (n = 100 & n = 500). Further investigation using real dataset from the UCI repository (Bupa Liver (n = 345) and Diabetes Messidor, n = 1151)) confirmed the imbalanced ratio effect on the parameter estimates and the odds ratio, and thus will lead to misleading results.

Keywords: Imbalance data · Parameter estimates · Logistic regression · Simulation, predictive analytics

1 Introduction

Classification problems with class imbalance, whereby one class has more observations than the other, emerge in many data mining applications, ranging from medical diagnostics [1–5], finance [6–8], marketing [9], manufacturing [10] and geology [11]. Due to their practical importance, the class imbalance problem have been widely studied by many researchers [12–21].

Logistic regression (LR) is a conventional statistical method and often used in predictive analytics as a benchmark when other classifiers are used. However, the imbalance situation creates a challenge for LR, whereby the focus of the many classifiers

M.W. Berry et al. (Eds.): SCDS 2016, CCIS 652, pp. 136–147, 2016.
DOI: 10.1007/978-981-10-2777-2_12

is normally on the overall optimization without taking into account the relative distribution between the classes [22]. Hence, the classification results obtained from the classifiers tends to be biased towards the majority class. Unfortunately, this imbalance problem is prominent as majority of the real dataset, regardless of field, suffers from some imbalance in nature [23]. This leads to the fact that the class with fewer observations is often misclassified into the majority classes [12, 14, 24].

Past studies either reported the effect of imbalance dataset on classifiers such as decision tree (DT), Support Vector Machine (SVM), and artificial neural network (ANN) [25–27] or uses LR on actual datasets to reveal the impact of IR in terms of predictions [2, 7, 19, 28]. However, no simulation study has been performed to determine the impact of imbalanced ratio on the estimate of the LR parameter (β) and thus the odds ratio (e^{β}) of the LR model.

The aim of this study is to investigate the effects of different IR on the parameter estimate of the binary LR model via a simulation study. This paper is organized as follows: Sect. 2 covers some previous studies on imbalanced data and applications of LR. The simulation methodology is explained in Sect. 3 and the results are presented in Sect. 4. Some discussions and the conclusion are given in Sect. 5.

2 Literature Review

Learning in the presence of imbalance data offers a great challenge to data scientists around the world. Techniques such as DT, ANN and SVM, performed well for balanced data but will not classify well when applied to imbalanced datasets [29]. [30] illustrated how overlapping within the dataset in the presence of imbalance data made it difficult for any classifier to predict observations correctly. A study by [31] demonstrated using actual datasets that the performance of a classifier (C4.5) is affected by the percentage of imbalance ratio.

Most studies reported the effect of IR towards the performance of standard classifiers using actual datasets with different ratio of imbalances [2, 17, 19, 25, 31]. [2] introduced REMED (Rule Extraction Medical Diagnosis), a 3-step algorithm using simple LR to select attributes for the model and adjust the percentage of the partition to improve accuracy of the model and evaluated it using real datasets. Although REMED is a competitive algorithm that can improve accuracy, it is restricted only to its domain, which is medical diagnostics. [17] reported that IR, dataset size and complexity, all contribute to the predictive performance of a classifier. They categorized the IR into three categories (balance, small, large), dataset size into four categories (very small, small, medium, and large) and complexity of the dataset into four categories (small, medium, large and very large) and experimented on different classifier (k-nearest neighbor (KNN), C4.5, SVM, multi-layered perceptron (MLP), naïve bayes (NB), and Adaboost (AB)). They concluded that higher IR has more influence on the performance of all the classifiers. [19] performed an extensive experiment on 35 dataset with different ratio of imbalance (1.33 %−34.90 %) using different sampling strategies (random oversampling (ROS), random undersampling (RUS), one-sided selection (OSS), cluster-based oversampling (CBOS), Wilson's editing (WE), SMOTE (SM), and borderline-SMOTE (BSM) on different classifiers (NB, DT C4.5, LR, random forest

(RF), and SVM). Their study significantly shows that sampling strategy improves the performance of the chosen classifiers and different classifiers works best with different sampling strategy. [25] observed that actual datasets do not provide the best distribution for any classifiers and sampling strategy is needed to improve the predictive performance of classifiers. Their experiment involves five actual dataset with C4.5 as the classifier. The study shows that SMOTE improve the performance of the overall classifier better than other sampling strategies and RUS works better than ROS with replication. [31] also experimented on 22 real dataset with different IR on different classifiers (C4.5, C4.5Rules, CN2 and RIPPER, Back-propagation Neural Network, NB and SVM). Using different sampling strategies (ROS, SMOTE, borderline-SMOTE, AdaSyn, and MetaCost), they concluded that the most affected in terms of accuracy (AUC) is the rule-based algorithm (C4.5Rule, RIPPER) and the least affected is the statistical algorithm (SVM).

From the studies mentioned above, we found that the performance of different standard classifiers such as NB, DT, ANN, and SVM were compared and each study concluded differently as to which classifier performed better [2, 17, 28, 32].

In predictive analytics, LR is an important classifier as it provides important information about the effect of an independent variable (IV) on the dependent variable (DV) through the odds ratio, which is the exponential value of the regression parameter [33]. LR is a statistical model and it is often used as a benchmark when other classifiers are used. However, with the presence of imbalance, the predictive models seemingly underestimate the class probabilities for minority class, despite evidently good overall calibration [34]. Significant odds ratio indicates a significant relationship between the DV and IVs.

The simulation study by [35] found that when sample size is large which is at least 500, the parameter estimates accuracy for LR improves. Their simulation study shows that the estimation of LR parameters is severely affected by types of covariates (continuous, categorical, and count data) and sample size. Meanwhile, [28] did a simulation study to evaluate the performance of six types of classifiers (ANN, Linear Discriminant Analysis (LDA), RF, SVM and penalized logistic regression (PLR)) on highly imbalanced data. However, their results show that the PLR with ROS method, failed to remove the biasness towards the majority class.

Most studies investigated the issue of IR by applying various classifiers to several datasets and thus there is no conclusive evidence as to which classifier is the best as it depends on the sample size, severity of imbalanced and types of data. Furthermore, no simulation study has been performed to determine the impact of IR on performance of classifiers. Simulation studies can provide empirical evidence on the impact of IR on the β-value and the odds ratio of the LR model.

3 Methods

There are two unknown parameters, β_0 and β_1 for a simple binary LR model. The parameters are estimated using the maximum likelihood method. The likelihood function by assuming observations to be independent is given by the following equation [33]:

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i} \tag{1}$$

The maximization of the likelihood function is required in order to estimate β_0 and β_1. Equivalently, the maximization of the natural logarithm of the likelihood function can be denoted by the following:

$$\log\left[L(\beta_{0,}\beta_1)\right] = \sum_{i=1}^{n} \{y_i \log[\pi(x_i)] + (1 - y_i) \log[1 - \pi(x_i)]\} \tag{2}$$

Referring to simple LR equation model, the equation in (2) can be expressed as follows [33]:

$$\log\left[L(\beta_{0,}\beta_1)\right] = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} \log[1 + \exp(\beta_0 + \beta_1 x_i)] \tag{3}$$

To maximize (3), one of the approaches is by differentiating $\log\left[L(\beta_{0,}\beta_1)\right]$ with respect to β_0 and β_1, and set the result of these two equations equal to zero as such:

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0 \text{ and } \sum_{i=1}^{n} x_i[y_i - \pi(x_i)] = 0 \tag{4}$$

One of the methods that can be used to solve Eq. (4) is using iterative computing methods. The maximum likelihood estimates of β_0 and β_1, are denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$. The maximum likelihood estimate of probability that event occurs, $\pi(x_i)$ is for case i denoted by:

$$\hat{\pi}(x_i) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}} \tag{5}$$

$\hat{\pi}(x_i)$ is also known as fitted or predicted value and the sum of $\hat{\pi}(x_i)$ is equal to the sum of the observed values:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}(x_i) \tag{6}$$

The estimated logit function is written as follows:

$$\hat{g}(x_i) = \log\left[\frac{\hat{\pi}(x_i)}{1 - \hat{\pi}(x_i)}\right] = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{7}$$

This study assesses the effect of different percentages of IR on estimation of parameter coefficients for binary LR model. The fitted model was compared with the true logistic model in order to assess the effect on the parameter estimation. The simulations were performed using R, an open source programming software.

We fit a binary LR model to the simulated data and obtain the estimated coefficients, $\hat{\beta}$. The study focuses on a model with a single continuous covariate. For a single covariate, the value of the regression coefficient (β_1) for the logistic model was set at 2.08 which gives a significant odds ratio (OR) of 8.004 for X ($OR = e^{2.08} = 8.004$).

We considered eight ratios: 1 %, 2 %, 5 %, 10 %, 20 %, 30 %, 40 %, and 50 % where IR 5 % or less represents severe imbalance in the response variable. However, the complexity of generating the simulated dataset specifically for fixed percentages of IR requires β_0 values to vary for different IR percentages. The full LR model is thus set as the following:

$$g(x) = \beta_{0k} + 2.08x_k \tag{8}$$

where β_{0k} is determined by the IR and k = 1, 2, 3, …. 8.

The distribution of the covariate (X) considered in this study is the standard normal distribution, *N(0,1)*. The sample sizes generated ranges from 100, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, and 5000. The simulation involves 10,000 replications. The R code developed for this simulation is made available at https://github.com/hezlin/simulationx1.git. It is also provided in the Appendix.

4 Simulation Results

Table 1 summarizes the results of parameter estimates for different sample sizes and IR. The results show that the estimates of β_0 and β_1 are far from the true parameter values for smaller sample size (n = 100, 500) especially with high IR value (IR = 1 %, 2 %, 5 %).

However, the estimates of the LR coefficient improve when the sample size increases to 1000 or more irrespective of IR. Figure 1 illustrates effect of IR and sample size. The red line is the set parameter value which is $\beta_1 = 2.08$.

Referring to Fig. 1, the effect on parameter estimate is most severe for small sample (n = 100). For sample size n = 500, at IR = 20 %, the value of the estimate is closer to the actual value. For sample size n = 1000 and n = 1500, at IR = 5 % the estimated β_1 is close enough to 2.08. Results show that for larger sample size (n = 1000 and above), the estimations are biased if the IR is less than 10 %.

The clustered boxplot in Fig. 2 shows clearly the effect of IR for various sample sizes. The estimates get closer to the true value when the sample size increases. The dispersion (standard deviation) of all $\hat{\beta}_1$ decreases as sample size and IR increases.

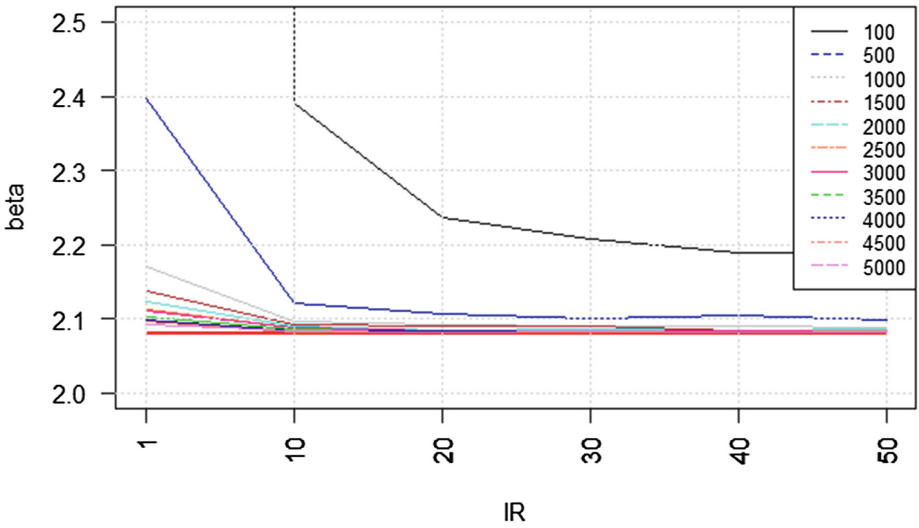

Fig. 1. Parameter estimates, $\hat{\beta}_1$, for different IR and sample size. This illustrates effect of different percentages of imbalance (IR = 1 %, 2 %, 5 %, 10 %, 20 %, 30 %, 40 %, 50 %) and sample size (n = 100, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000). The red line is the set parameter value, $\beta_1 = 2.08$. (Color figure online)

5 Application to Real Dataset

In this section, we illustrate the IR effect using two real datasets. The Bupa Liver and the Diabetes Messidor datasets were selected from the UCI repository. From the original dataset, we used stratified sampling to obtain the IR accordingly. The parameter estimates obtained are averaged over 1000 replications. Table 2 summarizes the results for this experiment.

The Bupa Liver Disorder dataset [36], consists of 345 observations, a response variable, *selector* (1 = *Yes* (58 %) and 0 = *No* (42 %)) and 6 covariates. We selected *alkaphos* (alkaline phosphate) as the independent variable as it is a continuous covariate. Results in Table 2 show that the value of $\hat{\beta}_0$ is far from the actual value as imbalance increases (that is when the percentage of IR gets smaller).

Although, the value of $\hat{\beta}_1$ and *OR* seems to be close enough to the true value from the original dataset, the p-value increases as IR increase (i.e. when imbalance decreases). Additionally, the confidence interval (CI) for $\hat{\beta}_1$ becomes wider for severe imbalance. For IR = 2 % and 1 %, the true value (-0.042) is no longer within the CI and thus the covariate will be conclude as not significant. These results show that IR affects the p-values for both estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$.

The Diabetes Messidor dataset [37], consist of 1151 observations. The response variable selected is *DR status* (1 = with DR (53 %) and 0 = without DR (47 %)) and

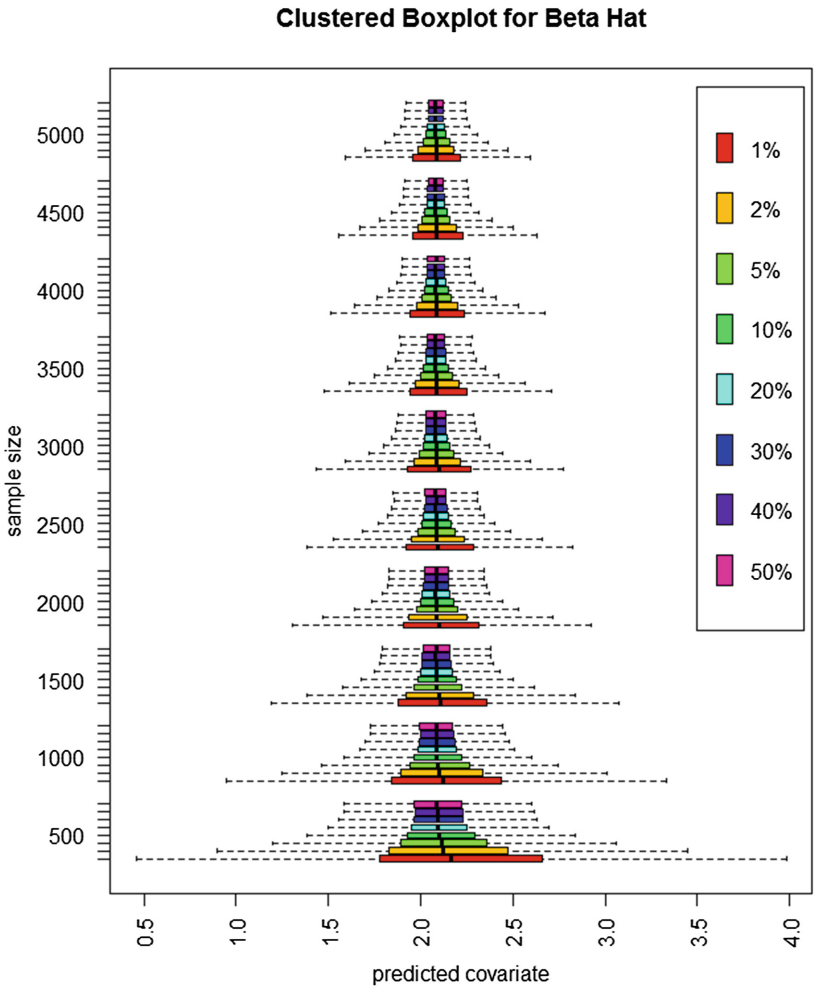

Fig. 2. Horizontal clustered boxplots for $\hat{\beta}_1$.

the dataset has 16 covariates. The covariate selected for this study is the *dmtroptdisc* (diameter of the optical disc). Results in Table 2 show that the $\hat{\beta}_0$ and $\hat{\beta}_1$ values becomes larger as the IR increases. The p-values for $\hat{\beta}_1$ increases as imbalance become more severe. The most obvious biased can be seen in the values of the odds-ratios which increase tremendously there is severe imbalance (at 5 % or less). These two applications confirm the results of the simulation study that imbalance data will lead to wrong conclusion on the effect of the independent variable on the response variable.

Table 1. Parameter Estimates for Different IR and Sample Size

Size	IR	$\hat{\beta}_1$	C.I (lower)	C.I (upper)	β_0	$\hat{\beta}_0$	Size	IR	$\hat{\beta}_1$	C.I (lower)	C.I (upper)	β_0	$\hat{\beta}_0$	Size	IR	$\hat{\beta}_1$	C.I (lower)	C.I (upper)	β_0	$\hat{\beta}_0$
100	**1**	82.173	94.183	70.163	-6.627	-193.893	**2000**	**1**	2.123	2.129	2.117	-6.565	-6.629	**4000**	**1**	2.098	2.102	2.094	-6.547	-6.575
	2	47.084	63.954	30.213	-5.764	-96.990		**2**	2.100	2.105	2.095	-5.726	-5.752		**2**	2.090	2.094	2.087	-5.710	-5.731
	5	4.822	6.061	3.582	-4.672	-9.117		**5**	2.093	2.096	2.090	-4.525	-4.541		**5**	2.088	2.091	2.086	-4.524	-4.534
	10	2.390	2.500	2.281	-3.566	-3.881		**10**	2.090	2.092	2.087	-3.498	-3.507		**10**	2.085	2.086	2.083	-3.496	-3.501
	20	2.236	2.247	2.224	-2.304	-2.403		**20**	2.084	2.087	2.082	-2.281	-2.284		**20**	2.084	2.085	2.082	-2.282	-2.285
	30	2.208	2.219	2.198	-1.427	-1.478		**30**	2.087	2.089	2.085	-1.417	-1.420		**30**	2.082	2.084	2.081	-1.417	-1.418
	40	2.190	2.199	2.180	-0.718	-0.720		**40**	2.085	2.087	2.083	-0.684	-0.685		**40**	2.082	2.084	2.081	-0.683	-0.684
	50	2.191	2.200	2.181	-0.274	-0.118		**50**	2.087	2.089	2.085	-0.002	0.000		**50**	2.083	2.084	2.082	0.000	0.000
500	**1**	2.398	2.518	2.278	-6.649	-7.259	**2500**	**1**	2.112	2.117	2.107	-6.558	-6.605	**4500**	**1**	2.097	2.101	2.093	-6.547	-6.573
	2	2.189	2.199	2.178	-5.760	-5.920		**2**	2.100	2.104	2.095	-5.726	-5.751		**2**	2.093	2.096	2.090	-5.718	-5.736
	5	2.146	2.153	2.139	-4.548	-4.621		**5**	2.089	2.091	2.086	-4.524	-4.535		**5**	2.086	2.088	2.084	-4.526	-4.531
	10	2.121	2.127	2.115	-3.505	-3.543		**10**	2.088	2.090	2.086	-3.498	-3.505		**10**	2.083	2.085	2.081	-3.496	-3.499
	20	2.106	2.111	2.102	-2.284	-2.303		**20**	2.085	2.087	2.083	-2.282	-2.285		**20**	2.082	2.083	2.081	-2.281	-2.283
	30	2.101	2.105	2.097	-1.420	-1.427		**30**	2.083	2.085	2.082	-1.418	-1.419		**30**	2.083	2.084	2.082	-1.418	-1.419
	40	2.104	2.108	2.100	-0.688	-0.691		**40**	2.084	2.085	2.082	-0.683	-0.684		**40**	2.082	2.083	2.081	-0.685	-0.685
	50	2.098	2.102	2.095	0.001	0.000		**50**	2.083	2.085	2.082	0.000	0.000		**50**	2.083	2.085	2.082	0.001	0.001
1000	**1**	2.170	2.179	2.160	-6.594	-6.729	**3000**	**1**	2.111	2.116	2.106	-6.552	-6.601	**5000**	**1**	2.092	2.096	2.089	-6.541	-6.564
	2	2.133	2.140	2.127	-5.814	-5.738		**2**	2.096	2.099	2.092	-5.720	-5.743		**2**	2.088	2.091	2.085	-5.713	-5.727
	5	2.110	2.115	2.105	-4.567	-4.529		**5**	2.088	2.091	2.086	-4.524	-4.535		**5**	2.087	2.089	2.085	-4.524	-4.533
	10	2.096	2.100	2.092	-3.501	-3.514		**10**	2.088	2.090	2.086	-3.498	-3.505		**10**	2.083	2.085	2.081	-3.497	-3.500
	20	2.092	2.096	2.089	-2.285	-2.292		**20**	2.085	2.087	2.083	-2.281	-2.284		**20**	2.081	2.082	2.080	-2.282	-2.282
	30	2.091	2.094	2.088	-1.418	-1.423		**30**	2.084	2.086	2.083	-1.419	-1.420		**30**	2.083	2.084	2.082	-1.418	-1.419
	40	2.091	2.093	2.088	-0.685	-0.687		**40**	2.084	2.085	2.082	-0.683	-0.684		**40**	2.082	2.083	2.080	-0.685	-0.685
	50	2.089	2.091	2.086	-0.002	-0.001		**50**	2.083	2.084	2.081	-0.048	-0.020		**50**	2.083	2.084	2.084	0.000	0.000
1500	**1**	2.138	2.145	2.131	-6.569	-6.662	**3500**	**1**	2.103	2.107	2.098	-6.551	-6.586							
	2	2.116	2.122	2.111	-5.731	-5.781		**2**	2.094	2.098	2.091	-5.712	-5.738							
	5	2.101	2.105	2.097	-4.528	-4.554		**5**	2.088	2.091	2.086	-4.525	-4.534							
	10	2.092	2.095	2.089	-3.499	-3.509		**10**	2.085	2.087	2.083	-3.497	-3.502							
	20	2.090	2.092	2.087	-2.283	-2.288		**20**	2.083	2.084	2.081	-2.281	-2.283							
	30	2.089	2.092	2.087	-1.417	-1.422		**30**	2.083	2.085	2.082	-1.417	-1.419							
	40	2.085	2.087	2.083	-0.684	-0.685		**40**	2.082	2.084	2.081	-0.685	-0.685							
	50	2.087	2.089	2.084	-0.001	0.000		**50**	2.083	2.084	2.081	0.000	0.000							

Table 2. Real dataset applications results

Dataset	Data/IR	$\hat{\beta}_0$, [p-value] C.I (lower, upper)	$\hat{\beta}_1$, p-value C.I (lower, upper)	*Odds-Ratio (OR)* C.I (lower, upper)
Bupa Liver (select status)	Original (145:200)	4.143, [0.069] (4.1433,4.1435)	−0.042, [0.093] (−0.04244, −0.042236)	0.959 (0.9584, 0.9585)
	40 % (133:200)	4.192, [0.077] (4.162, 4.221)	−0.042, [0.110] (−0.04226, −0.04161)	0.959 (0.9586, 0.9593)
	30 % (85:200)	4.530, [0.119] (4.454, 4.606)	−0.041, [0.205] (−0.042, −0.040)	0.960 (0.959, 0.961)
	20 % (50:200)	4.931, [0.179] (4.814, 5.048)	−0.039, [0.326] (−0.041, −0.038)	0.962 (0.960, 0.963)
	10 % (22:200)	5.773, [0.277] (5.573, 5.973)	−0.039, [0.447] (−0.042, −0.037)	0.962 (0.960, 0.964)
	5 % (10:200)	6.549, [0.361] (6.253, 6.845)	−0.039, [0.500] (−0.042, −0.036)	0.963 (0.960, 0.966)
	2 % (4:200)	7.436, [0.469] (6.987, 7.886)	−0.038, [0.554] (−0.043, −0.033)	0.965 (0.961, 0.970)
	1 % (2:200)	8.692, [0.503] (8.059, 9.325)	−0.044, [0.550] (−0.051, −0.037)	0.963 (0.957, 0.970)
Diabetes Messidor (DR status)	Original (540:611)	0.498, [0.170] (0.4975, 0.4977)	−3.449, [0.295] (−3.44852, −3.44854)	0.032 (0.031, 0.032)
	40 % (407:611)	0.771, [0.063] (0.766, 0.776)	−3.365, [0.372] (−3.411, −3.319)	0.079 (0.074, 0.085)
	30 % (261:611)	1.218, [0.019] (1.209, 1.227)	−3.381, [0.44] (−3.466, −3.297)	0.69 (0.419, 0.961)
	20 % (152:611)	1.752, [0.01] (1.738, 1.765)	−3.307, [0.482] (−3.431, −3.183)	104.45 (−52.948, 261.849)
	10 % (68:611)	2.514, [0.014] (2.491, 2.536)	−2.881, [0.511] (−3.089, −2.674)	116816.15 (218.528, 233413.773)
	5 % (35:611)	3.177, [0.025] (3.145, 3.209)	−2.821, [0.516] (−3.117, −2.524)	1.528E + 11 (−1.45E + 11, 4.51E + 11)
	2 % (13:611)	4.146, [0.069] (4.094, 4.199)	2.443, [0.517] (−2.929, −1.958)	1.535E + 20 (−1.47E + 20, 4.54E + 20)
	1 % (7:611)	4.664, [0.126] (4.593, 4.735)	−1.269, [0.513] (−1.926, −0.612)	8.043E + 26 (−6.63E + 26, 2.27E + 27)

6 Conclusion

This simulation study shows that the imbalance ratio in the response variable will affect the parameter estimates of the binary LR model. Imbalance ratio will lead to imprecise estimates and the bias is more severe if sample size is small. The imbalance ratio also affects the p-value of the parameter estimates and a covariate may be inaccurately reported as not significant. There are approaches suggested for handling imbalance data such as ROS, RUS, SMOTE, and clustering. Future work will include simulation study

involving categorical covariates and the classification performance of other classifiers such as SVM, DT, ANN and NB.

Acknowledgements. Our gratitude goes to the Research Management Institute (RMI) Universiti Teknologi MARA and the Ministry of Higher Education (MOHE) Malaysia for the funding of this research under the Malaysian Fundamental Research Grant, 600- RMI/FRGS 5/3 (16/2012). We also thank Prof. Dr. Haibo He (Rhodes Island University), Prof. Dr. Ronaldo Prati (Universidade Federal do ABC), Dr. Pam Davey and Dr. Carolle Birrell (University of Wollongong) for sharing their knowledge and providing valuable comments for this study.

Appendix

```
#fitting the model
set.seed(54321)
while(n<=nrep)
{ #set bnot value
  for(i in seq(start,end,0.0001))
  { x1 <- rnorm(ndata,0,1)      # some continuous variables
    z <- (i + 2.08*x1)   # linear combination with a bias
    pr1 <- 1/(1+exp(-z))
# pass through an inv-logit function
    ry<-runif(ndata,0,1) #generate value of y
    u <-as.matrix(ry) #generate value of y
    y <- ifelse((u<=pr1),1,0)
    y<-cbind(y)
    m_y <- (mean(y)*100)
    if(m_y == perc && n <=nrep)
    {dt<-data.frame(y=y,x=x1)
      mod<-glm(y~x1,data=dt,family="binomial")
#fit binary logistic regression
      beta0Hat[n]<-as.numeric(mod$coef[1])
      beta1Hat[n]<-as.numeric(mod$coef[2]) #store coefficient value
      resulty[n]<-as.numeric(sum(y))
      betanot[n]<-as.numeric(i)
      n <- n + 1 }
  }
}
mean(beta1Hat)
ci.b1 <- CI(beta1Hat,ci=0.95)
MSEbeta1Hat <- round(sum((beta1Hat-2.08)^2/nrep),3)
```

References

1. Datir, A.A., Wadhe, A.P.: Review on need of data mining techniques for biomedical field. Int. J. Comput. Inf. Technol. Bioinforma. **2**, 1–5 (2014)
2. Mena, L., Gonzalez, J.A.: Machine learning for imbalanced datasets: application in medical diagnostic. In: Proceedings of the Nineteenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2006), pp. 574–579. AAAI Press (2006). http://www.informatik.uni-trier.de/~ley/db/conf/flairs/flairs2006.html

3. Oztekin, A., Delen, D., Kong, Z.J.: Predicting the graft survival for heart-lung transplantation patients: an integrated data mining methodology. Int. J. Med. Inform. **78**, e84–e96 (2009)
4. Sathian, B.: Reporting dichotomous data using logistic regression in medical research: the scenario in developing countries. Nepal J. Epidemiol. **1**, 111–113 (2011)
5. Uyar, A., Bener, A., Ciray, H., Bahceci, M.: Handling the imbalance problem of IVF implantation prediction. IAENG Int. J. Comput. Sci. **37** (2010)
6. Akbani, R., Kwek, S.S., Japkowicz, N.: Applying support vector machines to imbalanced datasets. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) ECML 2004. LNCS (LNAI), vol. 3201, pp. 39–50. Springer, Heidelberg (2004)
7. Burez, J., Van den Poel, D.: Handling class imbalance in customer churn prediction. Expert Syst. Appl. **36**, 4626–4636 (2009)
8. Ogwueleka, F.: Data mining application in credit card fraud detection system. J. Eng. Sci. Technol. **6**, 311–322 (2011)
9. Nikulin, V., McLachlan, G.J.: Classification of imbalanced marketing data with balanced random sets. In: JLMR: Workshop and Conference Proceedings, vol. 7, pp. 89–100 (2009). http://jmlr.csail.mit.edu/proceedings/papers/v7/nikulin09/nikulin09.pdf
10. Sobran, N., Ahmad, A., Ibrahim, Z.: Classification of Imbalanced Dataset Using Conventional Naïve Bayes Classifier in 35–42 (2013). http://worldconferences.net/proceedings/aics2013/toc/papers_aics2013/A021-NURMAISARAHMOHDSOBRAN-ClassificationofImabalanceddatasetusingconventionalnaivebayesclassifier.pdf
11. Thogmartin, W.E., Knutson, M.G., Sauer, J.R.: Predicting regional abundance of rare grassland birds with a hierarchical spatial count model. Condor **108**, 25–46 (2006)
12. Chawla, N.V., Japkowicz, N., Kotcz, A.: Editorial: special issue on learning from imbalanced data sets. ACM SIGKDD Explor. Newsl. **6**, 1 (2004)
13. Drummond, C., Holte, R.: Severe class imbalance: why better algorithms aren't the answer. In: Gama, J., Camacho, R., Brazdil, P.B., Jorge, A.M., Torgo, L. (eds.) ECML 2005. LNCS (LNAI), vol. 3720, pp. 539–546. Springer, Heidelberg (2005)
14. He, H., Garcia, E.E.A.: Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. **21**, 1263–1284 (2009)
15. Japkowicz, N., Stephen, S.: The class imbalance problem: a systematic study. Intell. Data Anal. **6**, 429–449 (2002)
16. Japkowicz, N.: Learning from imbalanced data sets: a comparison of various strategies. In: AAAI Workshop on Learning from Imbalanced Data Sets 0–5 (2000). doi:10.1007/s13398-014-0173-7.2
17. Lemnaru, C., Potolea, R.: Imbalanced classification problems: systematic study, issues and best practices. In: Zhang, R., Zhang, J., Zhang, Z., Filipe, J., Cordeiro, J. (eds.) ICEIS 2011. LNBIP, vol. 102, pp. 35–50. Springer, Heidelberg (2012)
18. Longadge, R., Dongre, S.S., Malik, L.: Class imbalance problem in data mining review. Int. J. Comput. Sci. Netw. **2**, 83–87 (2013)
19. Van Hulse, J., Khoshgoftaar, T.M., Napolitano, A.: Experimental perspectives on learning from imbalanced data. In: Proceedings of 24th International Conference on Machine Learning, pp. 935–942 (2007). doi:10.1145/1273496.1273614
20. Visa, S., Ralescu, A.: Issues in mining imbalanced data sets - a review paper. In: Proceedings of the Sixteen Midwest Artificial Intelligence and Cognitive Science Conference, MAICS-2005, pp. 67–73 (2005)
21. Weiss, G.M.: Foundations of imbalanced learning. In: He, H., Ma, Y. (eds.) Imbalanced Learning: Foundations, Algorithms, Applications, pp. 13–42. Wiley & IEEE Press (2013). http://storm.cis.fordham.edu/gweiss/papers/foundations-imbalanced-13.pdf

22. Dong, Y., Guo, H., Zhi, W., Fan, M.: Class imbalance oriented logistic regression. In: 2014 International Conference Cyber-Enabled Distributed Computing and Knowledge Discovery, pp. 187–192 (2014). doi:10.1109/CyberC.2014.42
23. Goel, G., Maguire, L., Li, Y., McLoone, S.: Evaluation of sampling methods for learning from imbalanced data. Intell. Comput. Theor. **7995**, 392–401 (2013)
24. Weiss, G.M., Provost, F.: Learning when training data are costly: the effect of class distribution on tree induction. J. Arti. Intell. Res. **19**, 315–354 (2003)
25. Chawla, N.V.: C4. 5 and imbalanced data sets: investigating the effect of sampling method, probabilistic estimate, and decision tree structure. In: Proceedings of the International Conference on Machine Learning, Workshop Learning from Imbalanced Data Set II (2003). https://www3.nd.edu/∼dial/papers/ICML03.pdf
26. Cohen, G., Hilario, M., Sax, H., Hugonnet, S., Geissbuhler, A.: Learning from imbalanced data in surveillance of nosocomial infection. Artif. Intell. Med. **37**, 7–18 (2006)
27. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for the class imbalance problem bagging, boosting, and hybrid-based approaches. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **99**, 1–22 (2011)
28. Blagus, R., Lusa, L.: Class prediction for high-dimensional class-imbalanced data. BMC Bioinform. **11**, 523 (2010)
29. Anand, A., Pugalenthi, G., Fogel, G.B., Suganthan, P.N.: An approach for classification of highly imbalanced data using weighting and undersampling. Amino Acids **39**, 1385–1391 (2010)
30. Batista, G., Prati, R.C., Monard, M.C.: A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor. Newsl. **6**, 20 (2004)
31. Prati, R.C., Batista, G.E.A.P.A., Silva, D.F.: Class imbalance revisited: a new experimental setup to assess the performance of treatment methods. Knowl. Inf. Syst. 45, 247–270 (2014)
32. Sarmanova, A., Albayrak, S.: Alleviating class imbalance problem in data mining. In: Signal Processing and Communications Applications Conference, pp. 1–4 (2013)
33. Hosmer, D.W., Lemeshow, S.: Applied Logistic Regression Second Edition. Applied Logistic Regression (2004). doi:10.1002/0471722146
34. Wallace, B.C., Dahabreh, I.J.: Class Probability Estimates are Unreliable for Imbalanced Data (and How to Fix Them). *ICDM* (2012). http://www.cebm.brown.edu/static/papers/wallace-dahabreh-icdm-12-preprint.pdf
35. Hamid, H.A., Yap, B.W., Xie, X.-J., Abd Rahman, H.A.: Assessing the Effects of Different Types of Covariates for Binary Logistic Regression. **425,** 425–430 (2015)
36. Forsyth, R.S.: BUPA Liver Disorders (1990). https://archive.ics.uci.edu/ml/datasets/Liver+Disorders
37. Antal, B., Hajdu, A.: An ensemble-based system for automatic screening of diabetic retinopathy. Knowl. Based Syst. **60**, 20–27 (2014)