Research Article

Imbalanced Learning with Binary Regression: A Systematic Review and Some New Directions

Cheng Peng[a][1], Kai Peng[b]

[a]Department of Mathematics, West Chest University, West Chester, PA 19383, USA.
[b]School of Science, Ningbo University of Technology, Ningbo, Zhejiang 315048, China.

This work is in progress. Please do not cite or quote without the author's permission.

Contents

[1] Address of correspondence: C. Peng. Email: cpeng@wcupa.edu

ABSTRACT

Rare event detection (also called rare event mining or imbalanced learning) is a practically important but technically challenging problem. In the past two decades, various methods have been proposed to tackle this hard problem. These methods can be roughly classified into two major categories: algorithmic and modeling-based methods and sampling-based methods. All proposed methods have their own strengths and weaknesses as well. In this project, we focus on the models and algorithms that are rooted in the classical binary regression models and are suitable for rare event identification with stratified imbalanced learning. Some of the models are relatively new while others need to be further developed and investigated.

KEYWORDS

Machine learning; Logit model; Complementary log-log; Imbalanced data; Rare event mining; kernel regression; skewed links; non-parametric regression,

1. Introduction

Let y be a binary variable taking two possible values 1 (*success* class) and 0 (*failure* class), $\boldsymbol{x} = (x_1, x_2, \cdots, x_k)$ be a row vector of values of the k predictor variables, and $\boldsymbol{\theta} = (\beta_1, \beta_2, \cdots, \beta_k)^T$ be the slope parameters of the corresponding predictors. The binary logistic regression model has the following form

$$\ln\left(\frac{P(y=1|\boldsymbol{x})}{1-P(y=1|\boldsymbol{x})}\right) = \alpha + \boldsymbol{x}^T\boldsymbol{\beta} \tag{1}$$

The above logistic regression model is explicitly expressed in the following probability function

$$\pi(\boldsymbol{x}) = P(y=1|\boldsymbol{x}) = \frac{\exp(\alpha + \boldsymbol{x}^T\boldsymbol{\beta})}{1+\exp(\alpha + \boldsymbol{x}^T\boldsymbol{\beta})}. \tag{2}$$

When the above model is used as a predictive model, both intercept and slope parameters must be correctly estimated in order to estimate the *success* probability. The issue is when the classes are imbalanced, e.g., the *success* class is very rare compared to the *failure* class, the regular MLE of regression coefficients, particularly the intercept, are biased. This has been discussed in several studies. See, for example, the numerical examples in Owen [33] and Li et al. [28]. The imbalanced class bias results in the poor predictive power of models and algorithms.

In the past two decades, researchers have developed various logistic models and related algorithms to tackle imbalanced prediction problems. Among these researches, the work of Firth [12] and King and Zeng [21] have been the routine procedures of model-based imbalanced class prediction. Recently, numerous researchers modified and extended the work of Firth [12] and King and Zeng [21].

In Firth's approach, a penalized term was added to the likelihood function to remove the first order of bias in the MLE of the ordinary logistic model so that the resulting penalized estimators of the regression coefficient are less biased than the corresponding non-penalized estimators. This is a proactive approach to reducing the bias in the estimated regression coefficients. We will call the predictive logistic regression models based on Firth's penalized estimators the *Firth logistic model*. In the generalized linear model (exponential family with the classic link), Firth's added penalty function is equivalent to a non-informative Jeffery's [19] prior. In this sense, Firth logistic regression has some Bayesian flavor although the prior does not have a clear interpretation in the context of regression (see Gelman et al. [13]). As a general bias correction method, Firth logistic regression could be helpful in two different but related situations: imbalanced classes and complete/quasi-complete separation. Several papers demonstrated how to use the Firths method to handle separation in the logistic model (see, for example, Heinze and Schemper [16], Wang [46], and Zorn [49]).

Unlike Firth's preventive approach of bias reduction, King and Zeng [21] proposed to obtain the ordinary MLE of the regression coefficients of the logistic regression model and then modify the MLE of the intercept by including a Bayesian bias correction term. This correction term uses both prior information on the population class ratio and sample class ratio. We will refer to this bias correct model *King logistic Model*. We will show that King's bias corrective logistic model is equivalent to the retrospective case-control logistic regression model.

Qin [39] developed a semiparametric logistic regression model based on biased samples using Owen [32] empirical likelihood theory and the theory developed by Qin and Lawless [38]. We call this semiparametric logistic model *Qin logistic model*. The empirical likelihood function of the Qin logistic model is defined based on a stratified sample which allows taking two separate random samples independently from the two classes respectively. The resulting empirical likelihood estimator of the intercept involves prior information of the population class ratio that is also part of the bias correction term in King and Zeng [21]. When the prior imbalanced class ratio is known, the Qin logistic model can be naturally used for imbalanced class prediction without any penalization or correction.

Recently, several penalized methods including extensions and modifications of Firth and King models were also proposed in the literature to improve the prediction with imbalanced classes. To name a few, see the work of Kosmidis and Firth [24], Kosmidis and Firth [23], Kosmidis [22], Puhr et al. [37], Bergtold et al. [2], and Greenland and Mansournia [14]. Zhang et al. [48] proposed new penalized methods considering the cost of misspecification. As a more general form of penalization, various weighting methods have been studied to improve the estimator of model parameters. King and Zeng [21] proposed a simple weighting method using the imbalanced class ratios at both population and sample levels in the King logistic model. More non-intuitive methods based on boosting and adaptive boosting algorithms were also proposed by He and Cheng [15] and suggested to define the weighted likelihood of the logistic regression model. Motivated by Firth's penalized bias reduction and various modifications and extensions, Zhang [47] discussed various bias reduction methods for the Qin logistic model.

Various comparative studies regarding logistic regression models including both Firth and King models were conducted recently in the literature. See for example, Olmuş et al. [31], Faghih et al. [9], Bergtold et al. [2], Pavlou et al. [34], Rahman and Sultana [41], and among others.

Many performance metrics of the predictive ability of classification models and algorithms were discussed in the literature and implemented in real-world applications. Different applied fields have different preferred metrics that have practical interpretations in different fields. However, in rare event prediction, some of the well-known metrics such as accuracy, sensitivity, and specificity do not work well. Performance metrics that are more appropriate for rare event prediction have been discussed in the literature and used in practice (see, for example, Pinker [35], Leisman [27], Saito and Rehmsmeier [42], and Adhikari et al. [1], etc., and among others. How to find the optimal cut-off probability to calculate performance metrics was also discussed in the literature recently (see Boughorbel et al. [3],Calabrese [4]. Qin and Zhang [40] proposed a semiparametric estimation of the ROC curve using the Qin model. This method could also be used to estimate other global performance measures of the logistic model including the aforementioned measures for rare event prediction.

The primary objective of this study is to compare the predictive ability of three logistic models for imbalanced class data: the Firth penalized model, the King bias correct model, and the Qin semiparametric model based on biased samples. The optimal performance metrics are dependent on the model and the way of finding the best probability threshold that is dependent on the underlying performance metrics. Recently, Zou et al. [50] discussed the best way to find the best threshold in rare event classification.

The organization of this comparative study is as follows. In Section 2, we summarize the three aforementioned models from a methodological perspective to facilitate the numerical simulation and numerical experiments. Section 3 focuses on the simulation study. For ease of illustration, we restrict the simulation to the case that involves a single continuous predictor variable. Numerical experiments using various publicly available data sets with different imbalanced class ratios will be used with different performance metrics. Finally, some discussions and recommendations will be provided in the last section.

2. Basic Logit Modela

Assume $\{y_i, x_{1i}, x_{2i}, \cdots, x_{ki}\}$ (for $i = 1, 2, \cdots, n$) be an IID random sample. Let $\boldsymbol{x}_i = (x_{1i}, x_{2i}, \cdots, x_{ki})^T$ the vector of values of covariates of i-th observation and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ be the vector of the regression coefficients. Denote $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^T)^T$ and $\boldsymbol{X}_i = (1, \boldsymbol{x}_i^T)^T$. The (*perspective*) likelihood function of model (1) is given by

$$\mathscr{L}_{Prosp}(\boldsymbol{\theta}) = \prod_{i=1}^{n} [\pi(\boldsymbol{\theta} : \boldsymbol{x_i})]^{y_i} [1 - \pi(\boldsymbol{\theta} : \boldsymbol{x_i})]^{1-y_i} \tag{3}$$

The corresponding log-likelihood function is

$$\ell_{Prosp}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left\{ y_i \ln[\pi(\boldsymbol{\theta} : \boldsymbol{x}_i)] + (1 - y_i) \ln[1 - \pi(\boldsymbol{\theta} : \boldsymbol{x}_i)] \right\}, \tag{4}$$

where

$$\pi(\boldsymbol{\theta} : \boldsymbol{x}_i) = \frac{\exp(\alpha + \boldsymbol{x}_i^T \boldsymbol{\theta})}{1 + \exp(\alpha + \boldsymbol{x}_i^T \boldsymbol{\theta})} \overset{\text{def}}{=\!=} \pi_i(\boldsymbol{\theta}). \tag{5}$$

The score function is given by

$$\mathscr{U}(\boldsymbol{\theta}) = \frac{\partial \ell_{Prosp}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} [y_i - \pi_i(\boldsymbol{\theta})] \boldsymbol{X}_i.$$

The Information matrix of $\boldsymbol{\theta}$ is given by

$$\mathbf{I}(\boldsymbol{\theta}) = -\frac{\partial^2 \ell_{Prosp}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} = \sum_{i=1}^{n} \boldsymbol{X}_i^T \boldsymbol{X}_i \pi_i(\boldsymbol{\theta})[1-\pi_i(\boldsymbol{\theta})] = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X},$$

where $\boldsymbol{X}$ is the *design matrix* and

$$\boldsymbol{W} = \text{diag}(\pi_1(\boldsymbol{\theta})[1-\pi_1(\boldsymbol{\theta})], \cdots, \pi_n(\boldsymbol{\theta})[1-\pi_n(\boldsymbol{\theta})]) \tag{6}$$

be the $n \times n$ diagonal matrix. The MLE of $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\boldsymbol{\beta}}^T)$, is the solution to $\mathscr{U}(\boldsymbol{\theta}) = 0$. The predictive probability for new $\boldsymbol{x}_{new} = (x_{1,new}, x_{2,new}, \cdots, x_{k,new})^T$ of the ordinary logistic regression is given by

$$\hat{\pi}_{new} = \pi(\hat{\boldsymbol{\theta}} : \boldsymbol{x}_{new}) = \frac{\exp(\hat{\alpha} + \boldsymbol{x}_{new}^T \hat{\boldsymbol{\beta}})}{1 + \exp(\hat{\alpha} + \boldsymbol{x}_{new}^T \hat{\boldsymbol{\beta}})}. \tag{7}$$

3. Firth Penalized Logistic Model

The regression coefficients of the Firth model were estimated by maximizing the following penalized log-likelihood function

$$l_{Firth}(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) + \log\sqrt{\det|\mathbf{I}(\boldsymbol{\theta})|} \tag{8}$$

where $l_{Firth}(\boldsymbol{\theta})$ is the log-likelihood function of the ordinary logistic regression specified in (4). Assuming full rank exponential family, the information matrix $\mathbf{I}(\boldsymbol{\theta})$ is always positive semi-definite (see Chen et al. [7] and the recent work of Kosmidis and Firth [25]). The score function that is used to find the penalized MLE of the Firth model is given explicitly by

$$\mathscr{U}_{Firth}(\boldsymbol{\theta}) = \mathscr{U}(\boldsymbol{\theta}) - \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{\xi}, \tag{9}$$

where $\boldsymbol{X}$ is the design matrix and $\boldsymbol{W}$ was defined in (6). $\boldsymbol{\xi}$ is the vector of the diagonal elements of the following hat matrix

$$\boldsymbol{H} = \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{X} (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W}^{\frac{1}{2}} \tag{10}$$

with

$$\boldsymbol{W}^{\frac{1}{2}} = \text{diag}(\sqrt{\pi_1(\boldsymbol{\theta})[1-\pi_1(\boldsymbol{\theta})]}, \cdots, \sqrt{\pi_n(\boldsymbol{\theta})[1-\pi_n(\boldsymbol{\theta})]}). \tag{11}$$

The Firth's penalized MLE of $\boldsymbol{\theta}$, denoted by $\boldsymbol{\theta}^*$, is the solution to $\mathscr{U}_{Persp}(\boldsymbol{\theta}) = 0$ and the predictive probability based on $\boldsymbol{x}_{new} = (x_{1,new}, x_{2,new}, \cdots, x_{k,new})^T$ is given by

$$\bar{\pi}_{new}^{Firth} = \pi(\bar{\boldsymbol{\theta}} : \boldsymbol{x}_{new}) = \frac{\exp(\bar{\alpha} + \boldsymbol{x}_{new}^T \bar{\boldsymbol{\beta}})}{1 + \exp(\bar{\alpha} + \boldsymbol{x}_{new}^T \bar{\boldsymbol{\beta}})}. \tag{12}$$

The algorithms for finding $\bar{\boldsymbol{\theta}}$ were discussed in Firth [10, 11] and Olmuş et al. [31].

4. King's Adjusted Logistic Model

King and Zeng [21] corrected the potential bias due to the imbalanced class by adding the following Bayesian correction factor to the MLE of the intercept $\hat{\alpha}$ only and the MLE of other log odds ration parameter unchanged

$$\alpha_{bc} = -\ln\left[\left(\frac{1-\pi_0}{\pi_0}\right)\left(\frac{\hat{p}}{1-\hat{p}}\right)\right] \tag{13}$$

where π_0 is the population proportion of "successes" and $\hat{p}$ is the sample proportion of "successes". Let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k)^T$ and $\hat{\alpha}$ be the MLE derived from $\mathscr{U}(\boldsymbol{\theta}) = 0$. Denote $\hat{\boldsymbol{\theta}}_{bc} = (\hat{\alpha} + \alpha_{bc}, \hat{\boldsymbol{\beta}}^T)^T$. The predictive probability of the King model based on the value of new covariates,$\boldsymbol{x}_{new} = (x_{1,new}, x_{2,new}, \cdots, x_{k,new})^T$, is given by

$$\hat{\pi}_{new}^{King} = \pi(\hat{\boldsymbol{\theta}}_{bc} : \boldsymbol{x}_{new}) = \frac{\exp(\hat{\alpha} + \alpha_{bc} + \boldsymbol{x}_{new}^T\hat{\boldsymbol{\beta}})}{1 + \exp(\hat{\alpha} + \alpha_{bc} + \boldsymbol{x}_{new}^T\hat{\boldsymbol{\beta}})}. \tag{14}$$

Next, we derive the correction factor of the King model directly by maximizing the likelihood of parametric logistic regression based on retrospective case-control samples described in Prentice and Pyke [36] and the monograph of Hosmer and Lemeshow [17]. Let s be the sampling inclusion indicator variable of a finite population. $s = 1$ implies the associated subject will be included in either the control or treatment sample.

Consider the following two random samples taken retrospectively. Let

$$\{(y_1, \boldsymbol{x}_1), \cdots, (y_{n_0}, \boldsymbol{x}_{n_0})\} \sim P(x|y = 1) \tag{15}$$

be the random sample taken from the sub-population of *cases*,

$$\{(y_{n_0+1}, \boldsymbol{x}_{n_0+1}), \cdots, (y_{n_0+n_1}, \boldsymbol{x}_{n_0+n_1})\} \sim P(x|y = 0) \tag{16}$$

be the random sample from the sub-population of *controls*. Denote $n = n_0 + n_1$. The full likelihood function of the logistic regression model based on the above case-control data (aka. *retrospective likelihood function*) is given by

$$\mathscr{L}_{retrosp}(\boldsymbol{\theta}) = \prod_{i=1}^{n_0} P[\boldsymbol{x}_i|(y_i = 1) \cap (s = 1)] \prod_{i=n_0+1}^{n_0+n_1} P[\boldsymbol{x}_i|(y_i = 0) \cap (s = 1)] \tag{17}$$

We use the same notation on page 207 of Hosmer and Lemeshow [17]. Let $\tau_0 = P[s = 1|(y = 0) \cap \boldsymbol{x}] = P[s = 1|y = 0]$ be the sampling probability of including a subject in the sub-population of *controls* and $\tau_1 = P[s = 1|(y = 1) \cap \boldsymbol{x}] = P[s = 1|y = 1]$ the sampling probability of including a subject in the sub-population of cases. Note that $P\left[(y = 0|s = 1\right]$ is the proportion of cases among the selected subjects (i.e., the random sample of cases and controls). In other words. for the given sizes of the sub-sample of cases (n_1) and the sub-sample of controls (n_0), $P[(y = 0|s = 1] = n_0/(n_0 + n_1)$. Similarly, $P[(y = 1|s = 1] = n_1/(n_0 + n_1)$. Using the Bayes Theorem, we have

$$\tau_1 = P[s = 1|y = 1] = \frac{P[(y = 1|s = 1)P(s = 1)}{P(y = 1)} = \frac{n_1}{n_0 + n_1}\frac{P(s = 1)}{P(y = 1)} \tag{18}$$

similarly,

$$\tau_0 = P[s=1|y=0] = \frac{n_0}{n_0+n_1}\frac{P(s=1)}{P(y=0)}. \tag{19}$$

Dividing (19) by (18), we have

$$\frac{\tau_1}{\tau_0} = \frac{n_1}{n_0}\frac{P(y=0)}{P(y=1)} = \frac{n_1}{n_0}\frac{1-\pi_0}{\pi_0}. \tag{20}$$

Using the Bayes Theorem, we see the following connection between the retrospective and prospective likelihood functions at a single data point.

$$P[\boldsymbol{x}|y\cap(s=1)] = P[y|\boldsymbol{x}\cap(s=1)] \times \frac{P[\boldsymbol{x}|s=1]}{P[y|s=1]}. \tag{21}$$

Applying Bayes rule and the definition (2) (see details in Hosmer and Lemeshow [17] or Huang and Pepe [18]), we have

$$P[y=1|\boldsymbol{x}\cap(s=1)] = \frac{\tau_1 P[y=1|\boldsymbol{x}]}{\tau_1 P[y=1|\boldsymbol{x}] + \tau_0 P[y=0|\boldsymbol{x}]} = \frac{\tau_1 \exp(\alpha+\boldsymbol{x}^T\boldsymbol{\beta})}{\tau_0 + \tau_1\exp(\alpha+\boldsymbol{x}^T\boldsymbol{\beta})} \tag{22}$$

Similarly,

$$P[y=0|\boldsymbol{x}\cap(s=1)] = \frac{\tau_0 P[y=0|\boldsymbol{x}]}{\tau_0 P[y=0|\boldsymbol{x}] + \tau_1 P[y=1|\boldsymbol{x}]} = \frac{\tau_0}{\tau_0 + \tau_1\exp(\alpha+\boldsymbol{x}^T\boldsymbol{\beta})}. \tag{23}$$

Therefore,

$$\frac{P[y=1|\boldsymbol{x}\cap(s=1)]}{P[y=0|\boldsymbol{x}\cap(s=1)]} = \frac{\tau_1}{\tau_0}\exp(\alpha+\boldsymbol{x}^T\boldsymbol{\beta}) = \exp[\alpha + \ln(\tau_1/\tau_0) + \boldsymbol{x}^T\boldsymbol{\beta}] \tag{24}$$

Denote $\gamma = \alpha + \ln(\tau_1/\tau_0) = \alpha$ and

$$\mathscr{L}_{prosp}(\gamma,\boldsymbol{\beta}) = \prod_{i=1}^{n_0} P[y_i=1|\boldsymbol{x}_i\cap(s=1)] \prod_{i=n_0+1}^{n_0+n_1} P[y_i=0|\boldsymbol{x}_i\cap(s=1)] \tag{25}$$

Therefore, we have the following relationship between the case-control and prospective likelihood function

$$\mathscr{L}_{retrosp}(\boldsymbol{\theta}) = \mathscr{L}_{prosp}(\gamma,\boldsymbol{\beta}) \left(\prod_{i=1}^{n} \frac{P[\boldsymbol{x}_i|s=1]}{P[y_i|s=1]}\right) \tag{26}$$

The last term of the above equation is independent of parameters. Therefore, $\boldsymbol{\phi} = (\gamma, \boldsymbol{\beta}^T)$ can be estimated by

$$\widehat{\boldsymbol{\phi}} = \max_{\gamma,\boldsymbol{\beta}} \mathscr{L}_{prosp}(\gamma,\boldsymbol{\beta}). \tag{27}$$

This means we can fit model (1) to the case-control data (15, 16) as it was prospectively collected to obtain the MLE $(\widehat{\gamma}, \widehat{\boldsymbol{\beta}})$. The intercept α for predicting success probability based on model (1) is estimated by

$$\widehat{\alpha} = \widehat{\gamma} - \ln\left[\left(\frac{1-\pi_0}{\pi_0}\right)\left(\frac{\hat{p}}{1-\hat{p}}\right)\right]. \tag{28}$$

Hence, we have derived King and Zeng's correction factor from the retrospective likelihood function. We call $(\widehat{\alpha}, \widehat{\boldsymbol{\beta}})$ obtained above *retrspective parametric maximum likelihood estimate* (RPMLE) of $(\alpha, \boldsymbol{\beta})$ in model (1).

5. Qin's Semiparametric Model

We have derived the bias correction factor of the King model from maximizing the likelihood function of the retrospective logistic regression model in Section 4. In this section, we outline Qin's semiparametric retrospective logistic model using the empirical likelihood theory of Owen [32]. Assuming the same retrospective samples in (15) and (16). To be more specific, assume that

$$\begin{aligned} \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{n_0} &\text{ are i.i.d. from } P(x|y=1)\text{with density } g(\boldsymbol{x}), \\ \boldsymbol{x}_{n_0+1}, \boldsymbol{x}_{n_1+2}, \cdots, \boldsymbol{x}_{n_0+n_1} &\text{ are i.i.d. from } P(x|y=0)\text{with density } f(\boldsymbol{x}). \end{aligned} \tag{29}$$

From (21), we have

$$\frac{P[\boldsymbol{x}|y=1\cap(s=1)]}{P[\boldsymbol{x}|y=0\cap(s=1)]} = \frac{P[y=1|\boldsymbol{x}\cap(s=1)]}{P[y=0|\boldsymbol{x}\cap(s=1)]}\frac{P[y=0|s=1]}{P[y=1|s=1]}$$

$$= \exp\left[\alpha - \ln\left(\frac{1-\pi_0}{\pi_0}\frac{n_1}{n_0}\right) + \boldsymbol{x}^T\boldsymbol{\beta}\right]\frac{n_1}{n_0} = \exp\left[\alpha - \ln\left(\frac{1-\pi_0}{\pi_0}\right) + \boldsymbol{x}^T\boldsymbol{\beta}\right]. \tag{30}$$

This implies that

$$g(x) = \exp(\delta + \boldsymbol{x}^T\boldsymbol{\beta})f(x), \tag{31}$$

where $\delta = \alpha - \ln[(1-\pi_0)/\pi_0]$. Let $F(x)$ and $G(x)$ be the CDF of $f(x)$ and $g(x)$ respectively. Let $p_i = dG(t_i)$ $(i = 1, 2, \cdots, n)$ and $w(x) = \exp(\delta + \boldsymbol{x}^T\boldsymbol{\beta})$. The full semiparametric likelihood function under the above model (29) is defined to be

$$\mathscr{L}_{semipar}(\delta, \boldsymbol{\beta}, G) = \prod_{i=1}^{n} p_i \prod_{i=n_0+1}^{n_0} w(\boldsymbol{x}_i), \tag{32}$$

subject to the following constraints

$$\sum_{i=1}^{n} p_i = 1, p_i \geq 0, \sum_{i=1}^{n} p_i[w(\boldsymbol{x}_i) - 1] = 0 \tag{33}$$

For fixed δ and $\boldsymbol{\beta}$, $\mathscr{L}_{semipar}(\delta, \boldsymbol{\beta}, G)$ is maximized at

$$p_i = \frac{1}{n_0\{1 + \rho \exp[\delta + \boldsymbol{x}^T\boldsymbol{\beta}]\}}, \tag{34}$$

where $\rho = n_0/n_1$. Plugging the above p_i into the full semiparametric likelihood (32) to obtain the *profile likelihood* function. The retrospective semiparametric maximum likelihood estimates (RSMLE) of $(\delta, \boldsymbol{\beta})$, denoted by $(\widetilde{\delta}, \widetilde{\boldsymbol{\beta}})$, maximizes the following kernel of the profile log-likelihood function

$$l_{profile}(\delta, \boldsymbol{\beta}) = \sum_{i=1}^{n}[1 + \rho \exp(\delta + \boldsymbol{x}_i^T\boldsymbol{\beta})] + \sum_{i=n_0+1}^{n_0} \exp(\delta + \boldsymbol{x}_i^T\boldsymbol{\beta}). \tag{35}$$

In other words, the semiparametric maximum empirical likelihood estimate $(\tilde{\delta}, \tilde{\boldsymbol{\beta}})$ is the solution to the following system score equations

$$\frac{\partial l_{profile}(\delta, \boldsymbol{\beta})}{\partial \delta} = n_1 - \sum_{i=1}^{n} \frac{\rho \exp(\delta + \boldsymbol{x}_i^T\boldsymbol{\beta})}{1 + \rho \exp(\delta + \boldsymbol{x}_i^T\boldsymbol{\beta})} = 0 \tag{36}$$

$$\frac{\partial l_{profile}(\delta, \boldsymbol{\beta})}{\partial \delta} = \sum_{i=n_0+1}^{n_0+n_1} \boldsymbol{x}_i - \sum_{i=1}^{n} \frac{\boldsymbol{x}_i \rho \exp(\delta + \boldsymbol{x}_i^T\boldsymbol{\beta})}{1 + \rho \exp(\delta + \boldsymbol{x}_i^T\boldsymbol{\beta})} = 0 \tag{37}$$

The semiparametric estimate of the intercept α in model (1) given by $\widetilde{\alpha} = \widetilde{\delta} - \ln(\pi_1/\pi_0)$. Therefore, when the prior information of π_0 and π_1 are available, the predictive probability under the Qin model with new values of the covariate $\boldsymbol{x}_{new} = (x_{1,new}, x_{2,new}, \cdots, x_{k,new})^T$, is given by

$$\widetilde{\pi}_{new}^{Qin} = \frac{\exp[\widetilde{\alpha} + \boldsymbol{x}_{new}^T\widetilde{\boldsymbol{\beta}}]}{1 + \exp[\widetilde{\alpha} + \boldsymbol{x}_{new}^T\widetilde{\boldsymbol{\beta}}]}. \tag{38}$$

6. Connection between King and Qin Models

For ease of comparison, we use the ordinary logistic regression model as a reference. The MLE of the ordinary logistic regression coefficients is the solution of the prospective likelihood of the model (1) given in (4).

$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = \max_{\alpha, \boldsymbol{\beta}} \ell_{Prosp}(\alpha, \boldsymbol{\beta}) \tag{39}$$

The penalized maximum likelihood estimate (PMLE) of the coefficients of the Firth model is the solution to the following optimization problem related to (8)

$$(\overline{\alpha}, \overline{\boldsymbol{\beta}}) = \arg\max_{\alpha, \boldsymbol{\beta}} \left(\ell(\alpha, \boldsymbol{\beta}) + \log \sqrt{det|\boldsymbol{I}(\alpha, \boldsymbol{\beta})|} \right) \tag{40}$$

Clearly, the MLE and PMLE of the regression coefficients are different. The Firth model is a member of a larger family of regularized logistic regression models that includes ridge logistic regression and LASSO (Least Absolute Shrinkage and Selection Operator) logistic regression. Ridge and LASSO logistic models are also used to improve the predictive capability. More information and applications of these regularized logistic regression models can be found in Le Cessie and Van Houwelingen [26], Månsson and Shukur [29], Schaefer et al. [43], Kibria et al. [20], Meier et al. [30], Meier et al. [30], Wang et al. [45], etc.

The King and Qin models are members of the family of retrospective case-control logistic regression models. The RPMLE of coefficients of King's bias-corrected model are estimated by maximizing the following prospective likelihood function.

$$(\widehat{\gamma}, \widehat{\boldsymbol{\beta}}) = \arg\max_{\gamma, \boldsymbol{\beta}} \mathscr{L}_{Prosp}(\gamma, \boldsymbol{\beta}). \tag{41}$$

The RPMLE of $\boldsymbol{\beta}$ obtained in (41) is identical to the MLE obtained from (39). The RPMLE and MLE of the intercept are different. The RSMLE of the coefficients of the Qin model maximizes the semiparametric profile log-likelihood function (35) as follows

$$(\widetilde{\delta}, \widetilde{\boldsymbol{\beta}}) = \arg\max_{\delta, \boldsymbol{\beta}} \ell_{profile}(\delta, \boldsymbol{\beta}) \tag{42}$$

The RSMLE of the intercept of model (1) is given by $\widetilde{\alpha} = \widetilde{\delta} - \log(\pi_1/\pi_0)$. The term $\log(\pi_1/\pi_0)$ can be considered as the bias correction factor that is similar to that in the King model. Note that the odds ratio parameters are dependent on the ratio of the sizes of the subsamples of cases and controls. However, if both sub-sample sizes are equal (i.e., $n_0 = n_1$), the King model is identical to the Qin model.

We have shown that the King and Qin models represent parametric and semiparametric retrospective case-control logistic regression models. We could also consider other two semiparametric case-control logistic regression models that have the potential to improve the predictive capacity of the imbalanced class.

7. Binary Kernel Logistic Regression

The binary kernel logistic regression introduces the kernel function to tackle nonlinear systematic components in the standard logistic regression model.

Let the i-th row vector $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \cdots, x_{i,(p-1)}, x_{ip})$ of the following design matrix.

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p-1} & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,p-1} & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n-1,1} & x_{n-2,2} & \cdots & x_{n-1,p-1} & x_{n-1,p} \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,p-1} & x_{np} \end{bmatrix}$$

To form the general kernel logistic regression model, we consider the simplest form of an identity mapping polynomial basis function 0 of the feature space such that

$$\phi(\mathbf{x}_i) = \phi[(x_{i1}, x_{i2}, \cdots, x_{i,(p-1)}, x_{ip})] = \mathbf{x}_i.$$

The logit link function can be rewritten as

$$\eta_i = \beta_0 + \phi(\mathbf{x_i})\beta^T$$

In general, function $\phi(\cdot)$, maps the data from the lower dimension space to the higher dimension space, so that

$$\phi : \mathbf{x} \in \mathbf{R}^d \rightarrow \phi(\mathbf{x}) \in \mathbf{F}$$

The kernel logistic regression is one of the best-known machine-learning techniques for classification and has been applied widely (see Tien Bui et al. [44], Cawley and Talbot [6], Cawley and Talbot [5], etc.). To estimate the class-posterior probabilities with the kernel's log-linear function combination by applying the penalized maximum likelihood method. The nonlinear form of the logistic can be formulated as follows

$$\text{logit} P(Y = 1|x) = \omega\phi(x) + b \tag{43}$$

where *omega* and α are the optimal model parameters obtained by minimizing a cost function, which represents the regularized negative-log likelihood of the data.

8. Friedman Additive Logistic Model

Friedman's additive logistic Model is a non-linear non-parametric logistic regression model that is defined by

$$\log \frac{P(Y = 1|x)}{P(Y = 0|x)} = \alpha + f_1(x1) + f_2(x_2) + \cdots + f_k(x_k).$$

Each predictor x_i enters the model individually through adding function $f_i(x_i)$. No interaction terms such as $f(x_1, x_2)$ are used to reflect the interaction. Apparently, the additive logistic regression model is reduced to the standard logistic regression model when $f_i(x_i) = \beta_i x_i$.

The additive is nonlinear since $f_i(x_i)$ is a nonlinear function. It is also a non-parametric model since $f_i(x_i)$ is not specified and it is estimated based on the given data using, for example, non-parametric spline methods.

How to adjust the model to make it suitable for rare event learning from imbalanced data has been explored in the literature. We could borrow the idea of King [21] and Qin [39] to build an adjusted additive logistic model for an imbalanced learning algorithm.

9. Skewed Logit Models for Imbalanced Learning

All previously surveyed or could-be-modified binary regression models for potential imbalanced learning are based on the symmetric logit link function. The symmetric link functions approach zero and one at the same speed. Huayanay et al de la

Cruz Huayanay et al. [8] performed a simulation study comparing the RMSE of several logistic regression models and binary regression with skewed links. The results show that models with skewed links outperformed the other logistic regression models.

10. Parametric and Non-parametric C-Loglog and Loglog Models

All previously surveyed or could-be-modified binary regression models for potential imbalanced learning are based on the symmetric logit link function. The symmetric link functions approach zero and one at the same speed.

The log-log link function is defined as $-\log[-\log[E(Y)]]$. Note that using the log-log link function for the probability of "successes" is the same as using the complementary log-log link function for the probability of "failures" in a generalized linear model. The coefficients from the two models only differ by sign. Some authors define the log-log link function as $\log[-\log[E(Y)]$ and the inverse link function is $\exp[-\exp(\eta)]$ so that in a GLM the signs of the parameters are reversed and are equal to the parameters of a model with a complementary log-log link function for the probabilities of failures. We only focus on the C-loglog model.

Parametric C-loglog models have been used on various domain-specific applications. Because of the structure of the model, it

11. Potential New Semiparametric Skewed Binary Regression Models

Both kernel and skewed link-based logistic models discussed in the previous sections can be modified to semiparametric models to use stratified samples so that can be used for imbalanced data for rare event detection. We will present several proposals in this section.

12. Comparison of Models

The work of Huayanay et al de la Cruz Huayanay et al. [8] showed the better performance of the skewed model in terms of the standard errors of the estimated regression coefficient. In this section, we perform a simulation study to compare several models such as Qin's semiparametric logit, semiparametric C-loglog, and other adjusted and penalized logistic regression models in terms of predictive power with an imbalanced design.

References

[1] Samrachana Adhikari, Sharon-Lise Normand, Jordan Bloom, David Shahian, and Sherri Rose. Revisiting performance metrics for prediction with rare outcomes. *Statistical Methods in Medical Research*, 30(10):2352–2366, 2021.

[2] Jason S Bergtold, Elizabeth A Yeager, and Allen M Featherstone. Inferences from logistic regression models in the presence of small samples, rare events, nonlinearity, and multicollinearity with observational data. *Journal of Applied Statistics*, 45(3):528–546, 2018.

[3] Sabri Boughorbel, Fethi Jarray, and Mohammed El-Anbari. Optimal classifier for imbalanced data using matthews correlation coefficient metric. *PloS one*, 12(6): e0177678, 2017.

[4] Raffaella Calabrese. Optimal cut-off for rare events and unbalanced misclassification costs. *Journal of Applied Statistics*, 41(8):1678–1693, 2014.

[5] Gavin C Cawley and Nicola LC Talbot. Efficient model selection for kernel logistic regression. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 2, pages 439–442. IEEE, 2004.

[6] Gavin C Cawley and Nicola LC Talbot. Efficient approximate leave-one-out cross-validation for kernel logistic regression. *Machine Learning*, 71(2):243–264, 2008.

[7] Ming-Hui Chen, Joseph G Ibrahim, and Sungduk Kim. Properties and implementation of jeffreys's prior in binomial regression models. *Journal of the American Statistical Association*, 103(484):1659–1664, 2008.

[8] Alex de la Cruz Huayanay, Jorge L Bazan, Vicente G Cancho, and Dipak K Dey. Performance of asymmetric links and correction methods for imbalanced data in binary regression. *Journal of Statistical Computation and Simulation*, 89(9): 1694–1714, 2019.

[9] Marjan Faghih, Zahra Bagheri, Dejan Stevanovic, Seyyed Mohhamad Taghi Ayatollahi, and Peyman Jafari. A comparative study of the bias correction methods for differential item functioning analysis in logistic regression with rare events data. *BioMed Research International*, 2020, 2020.

[10] David Firth. Bias reduction, the jeffreys prior and glim. In *Advances in GLIM and Statistical Modelling*, pages 91–100. Springer, 1992.

[11] David Firth. Generalized linear models and jeffreys priors: an iterative weighted least-squares approach. In *Computational statistics*, pages 553–557. Springer, 1992.

[12] David Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1): 27–38, 1993.

[13] Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. A weakly informative default prior distribution for logistic and other regression models. *The annals of applied statistics*, 2(4):1360–1383, 2008.

[14] Sander Greenland and Mohammad Ali Mansournia. Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Statistics in medicine*, 34(23):3133–3143, 2015.

[15] Jia He and Maggie X Cheng. Weighting methods for rare event identification from imbalanced datasets. *Frontiers in big Data*, 4:715320, 2021.

[16] Georg Heinze and Michael Schemper. A solution to the problem of separation in logistic regression. *Statistics in medicine*, 21(16):2409–2419, 2002.

[17] David W Hosmer and Stanley Lemeshow. Applied logistic regression. john wiley & sons. *New York*, 2000.

[18] Ying Huang and Margaret Sullivan Pepe. Assessing risk prediction models in case-control studies using semiparametric and nonparametric methods. *Statistics in medicine*, 29(13):1391–1410, 2010.

[19] Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

[20] BM Kibria, Kristofer Månsson, and Ghazi Shukur. Performance of some logistic ridge regression estimators. *Computational Economics*, 40(4):401–414, 2012.

[21] Gary King and Langche Zeng. Logistic regression in rare events data. *Political analysis*, 9(2):137–163, 2001.

[22] Ioannis Kosmidis. Bias in parametric estimation: reduction and useful side-effects. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(3):185–196, 2014.

[23] Ioannis Kosmidis and David Firth. Bias reduction in exponential family nonlinear models. *Biometrika*, 96(4):793–804, 2009.

[24] Ioannis Kosmidis and David Firth. A generic algorithm for reducing bias in parametric estimation. *Electronic Journal of Statistics*, 4:1097–1112, 2010.

[25] Ioannis Kosmidis and David Firth. Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika*, 108(1):71–82, 2021.

[26] Saskia Le Cessie and Johannes C Van Houwelingen. Ridge estimators in logistic regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(1):191–201, 1992.

[27] Daniel E Leisman. Rare events in the icu: an emerging challenge in classification and prediction. *Critical care medicine*, 46(3):418–424, 2018.

[28] Yazhe Li, Tony Bellotti, and Niall Adams. Issues using logistic regression with class imbalance, with a case study from credit risk modelling. *Foundations of Data Science*, 1(4):389, 2019.

[29] Kristofer Månsson and Ghazi Shukur. On ridge parameters in logistic regression. *Communications in Statistics-Theory and Methods*, 40(18):3366–3381, 2011.

[30] Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

[31] Hülya Olmuş, Ezgi Nazman, and Semra Erbaş. Comparison of penalized logistic regression models for rare event case. *Communications in Statistics-Simulation and Computation*, 51(4):1578–1590, 2022.

[32] Art B Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249, 1988.

[33] Art B Owen. Infinitely imbalanced logistic regression. *Journal of Machine Learning Research*, 8(4), 2007.

[34] Menelaos Pavlou, Gareth Ambler, Shaun Seaman, Maria De Iorio, and Rumana Z Omar. Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Statistics in medicine*, 35(7):1159–1177, 2016.

[35] Edieal Pinker. Reporting accuracy of rare event classifiers. *NPJ digital medicine*, 1(1):1–2, 2018.

[36] Ross L Prentice and Ronald Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411, 1979.

[37] Rainer Puhr, Georg Heinze, Mariana Nold, Lara Lusa, and Angelika Geroldinger. Firth's logistic regression with rare events: accurate effect estimates and predictions? *Statistics in medicine*, 36(14):2302–2317, 2017.

[38] Jin Qin and Jerry Lawless. Empirical likelihood and general estimating equations. *the Annals of Statistics*, 22(1):300–325, 1994.

[39] Jing Qin. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85(3):619–630, 1998.

[40] Jing Qin and Biao Zhang. Using logistic regression procedures for estimating receiver operating characteristic curves. *Biometrika*, 90(3):585–596, 2003.

[41] M Shafiqur Rahman and Mahbuba Sultana. Performance of firth-and logf-type penalized methods in risk prediction for small or sparse binary data. *BMC medical research methodology*, 17(1):1–15, 2017.

[42] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS*

one, 10(3):e0118432, 2015.

[43] RL Schaefer, LD Roi, and RA Wolfe. A ridge logistic estimator. *Communications in Statistics-Theory and Methods*, 13(1):99–113, 1984.

[44] Dieu Tien Bui, Tran Anh Tuan, Harald Klempe, Biswajeet Pradhan, and Inge Revhaug. Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. *Landslides*, 13(2):361–378, 2016.

[45] Hong Wang, Qingsong Xu, and Lifeng Zhou. Large unbalanced credit scoring using lasso-logistic regression ensemble. *PloS one*, 10(2):e0117844, 2015.

[46] Xuefeng Wang. Firth logistic regression for rare variant association tests, 2014.

[47] Biao Zhang. Bias-corrected maximum semiparametric likelihood estimation under logistic regression models based on case–control data. *Journal of statistical planning and inference*, 136(1):108–124, 2006.

[48] Lili Zhang, Trent Geisler, Herman Ray, and Ying Xie. Improving logistic regression on the imbalanced data by a novel penalized log-likelihood function. *Journal of Applied Statistics*, 49(13):3257–3277, 2022.

[49] Christopher Zorn. A solution to separation in binary response models. *Political Analysis*, 13(2):157–170, 2005.

[50] Quan Zou, Sifa Xie, Ziyu Lin, Meihong Wu, and Ying Ju. Finding the best classification threshold in imbalanced classification. *Big Data Research*, 5:2–8, 2016.