THE THEORETICAL STUDY OF RARE EVENT WEIGHTED LOGISTIC REGRESSION FOR CLASSIFICATION OF IMBALANCED DATA

Dian Eka Apriana Sulasih, Santi Wulan Purnami, Santi Puteri Rahayu

Institut Teknologi Sepuluh Nopember, Department of Statistics,
Jl. Arif Rahman Hakim, Surabaya 60111, Indonesia

dian.eka14@mhs.statistika.its.ac.id (Dian Eka A S)

Abstract

One of the problems in data classification is imbalanced data. In two-class classification, imbalance problem occurs where one of the two classes has more samples than another class. In such situation, most of the classifier will be biased towards the major class, while the minor class will be subordinated eventually which leads to inaccurate classification. Therefore, a method to classify the imbalanced data is required. Rare Event Weighted Logistic Regression (RE-WLR) which is developed by Maalouf and Siddiqi is a method of classification applied to large imbalanced data and rare event. This study showed the review of RE-WLR for the classification of imbalanced data. It explicated the steps to obtain the estimator specifically, particularly for IRLS. RE-WLR is a combination of Logistic Regression (LR) rare events corrections and Truncated Regularized Iteratively Re-weighted Least Squares (TR-IRLS). Rare event correction in LR is applied to Weighted Logistic Regression (WLR). Regularization was added to reduce over-fitting. The estimation of β is performed by using the method of *maximum likelihood* (ML), while WLR *maximum likelihood estimates* (MLE) were obtained by using IRLS method of Newton-Raphson algorithm. In order to solve large optimization problems, Truncated-Newton method is applied.

Keywords: two-class classification, imbalanced data, logistic regression, rare event, RE-WLR

Presenting Author's biography

Dian Eka Apriana Sulasih earned her bachelor's degree from Sekolah Tinggi Ilmu Statistik Jakarta in 2007. She is an government employee at Badan Pusat Statistik. Currently, she is also a post-graduate student of Statistics in Institut Teknologi Sepuluh Nopember.

1. Introduction

Data classification is an important process in the field of data mining. Classification is the processing of finding a set of models (or functions) which describe and distinguish data classes or concepts, for the purposes of being able to use the model to predict the class of objects whose class label is unknown [1]. The methods often used for classification are discriminant analysis and logistic regression. In addition, many classification methods are used by approach to programming, such as Artificial Neural Network (ANN), Naive Bayes, Adaptive Classification Regression Tree (CART) and Support Vector Machine (SVM).

Linear Discriminant Analysis and Logistic Regression methods search for linear functions which are subsequently used for classification purposes. The use of linear functions enables better interpretation of the results by analyzing the value of the coefficients obtained. Not every classification method permits this type of analysis and, in fact, some are classified as ‘‘black box” models. Hence, Classic Discriminant Analysis and Logistic Regression continue to be interesting methodologies [2].

Linear classification is an extremely important machine-learning and data-mining tool. Compared to other classification techniques, such as the Kernel methods, which transform data into higher dimensional space, linear classifiers are implemented directly on data in their original space. The main advantage of linear classifiers is their efficient training and testing procedures, especially when implemented on large and high-dimensional data sets [3].

Besides discriminant analysis, logistic regression is a method of classification with linear classifiers that is often used. Logistic regression, which is a linear classifier, has been proven to be a powerful classifier by providing probabilities and by extending to multi-class classification problems [4,5]. Logistic regression has been extensively studied. Moreover, LR requires solving only unconstrained optimization problems. Hence, with the right algorithms, the computation time can be much less than that of other methods, such as Support Vector Machines (SVM), which require solving a constrained quadratic optimization problem [6].

One of the problems in data classification is the composition of the data that is out of balance (imbalanced data). In the binary or two-class classification, one class has a greater number of samples than the other class. The majority is negative class while the minority is positive class. The problem that occurs is a good prediction accuracy of the negative class and poor prediction accuracy for positive class. In other words, classifier tends to predict class which has more data composition.

Problems of imbalanced data occur in various fields such as information retrieval and filtering [7], detection of oil spills from satellite imagery [8], medical diagnosis [9], text classification [10], credit card fraud detection [11], telecommunications [12], and others.

Most of algorithm is more preoccupied in classifying major sample and ignoring or misclassifying minor sample. The minor samples are those that rarely occur but very important. There are various methods available for classification of imbalance data set which is divided into three main categories, the algorithmic approach, data preprocessing approach, and feature selection approach. Each of this technique has their own advantages and disadvantages [13].

Some Logistic Regression method development has been done to improve the accuracy of classification in imbalanced data. Maalouf and Trafalis [14] developed a method of Rare Event Weighted Kernel Logistic Regression (RE-WKLR) that is suitable for small to medium-sized data. Rahayu [15] developed a method of AdaBoost Newton Truncated Regularized Weighted Kernel Logistic Regression (AB-WKLR) and AdaBoost NTR Weighted Regularized Logistic Regression (AB-WLR) that significantly increase the performance of the accuracy and stability of general classifiers at NTR-KLR and NTR-LR , Furthermore, Maalouf and Siddiqi [6] developed a method of Rare Event Weighted Logistic Regression (RE-WLR) for the classification of imbalanced data on large-scale data. The study concluded that the RE-WLR has better performance than Truncated-Regularized Iteratively Re-weighted Least Squares (TR-IRLS). Wang, Xu and Zhou [16] used Lasso-Logistic Regression to evaluate unbalanced credit scoring.

This study showed the review of RE-WLR developed by Maalouf and Siddiqi [6] for the classification of imbalanced data. It mainly explicated the steps to obtain estimator specifically, particularly for IRLS.

2. Logistic Regression

Let $\mathbf{X} \in R^{Nxd}$ be a data matrix where N is the number of instances (examples) and d is the number of features (parameters or attributes), and y be a binary outcomes vector. For every instance $\boldsymbol{x}_i \in R^d$ (a row vector in **X**), where i = 1 . . .N, the outcome is either $y_i = 1$ or $y_i = 0$. Let the instances with outcomes of $y_i = 1$ belong to the positive class occurrence of an event), and the instances with outcomes $y_i = 0$ belong to the negative class (non occurrence of an event). The goal is to classify the instance $\boldsymbol{x}_i$ as positive or negative. An instance can be treated as a Bernoulli trial with an expected value $E(y_i)$ or probability p_i. The logistic function commonly used to model each instance $\boldsymbol{x}_i$ with its expected outcome is given by the following formula [17]:

$$E(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}) = p_i = \frac{e^{\mathbf{x_i\beta}}}{1 + \mathrm{e}^{\mathbf{x_i\beta}}} \tag{1}$$

where β is the vector of parameters with the assumption that $x_{i0} = 1$ so that the intercept β_0 is a constant term. From then on, the assumption is that the intercept is included in the vector β.

The logistic (logit) transformation is the logarithm of the odds of the positive response and is defined as

$$\boldsymbol{\eta_i} = In\left(\frac{p_i}{1 - p_i}\right) = \mathbf{x_i\beta} \tag{2}$$

In matrix form, the logit function is expressed as

$$\boldsymbol{\eta} = \mathbf{X\beta} \tag{3}$$

3. Regularized Logistic Regression

There are two general methods of parameter estimation. They are least-squares estimation (LSE) and maximum likelihood estimation (MLE). This study uses Maximum Likelihood (ML). The ML method is based on the joint probability density of the observed data, and acts as a function of the unknown parameters in the model [18]. Recall that the outcome y is a Bernoulli random variable with mean p_i in the LR model. Therefore we may interpret the expectation function as the probability that $y = 1$, or equivalently that $\mathbf{x_i}$ belongs to the positive class. Thus we may compute the probability of the i-th experiment and outcome in the dataset **X**, **y** as

$$P(x_i, y_i)|\beta) = \begin{cases} p_i & ,\text{if } y = 1 \\ 1 - p_i & ,if\ y = 0 \end{cases} \tag{4}$$

$$= (p_i)^{y_i}(1 - p_i)^{(1-y_i)} \tag{5}$$

The likelihood function is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n}(p_i)^{y_i}(1 - p_i)^{(1-y_i)} = \prod_{i=1}^{n}\left(\frac{e^{\mathbf{x_i\beta}}}{1 + e^{\mathbf{x_i\beta}}}\right)^{y_i}\left(\frac{1}{1 + e^{\mathbf{x_i\beta}}}\right)^{(1-y_i)} \tag{6}$$

and hence, the log-likelihood is then

$$In\ L(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(y_i In\left(\frac{e^{\mathbf{x_i\beta}}}{1 + e^{\mathbf{x_i\beta}}}\right) + (1 - y_i)In\left(\frac{1}{1 + e^{\mathbf{x_i\beta}}}\right)\right) \tag{7}$$

There is no closed form solution to maximize $InL(\beta)$ with respect to β. Therefore, LR *maximum likelihood estimates* (MLE) are obtained using numerical optimization methods. One of the most commonly used numerical methods is the Newton-Raphson method, for which, both the gradient vector and the Hessian matrix are needed. The gradient vector is obtained from the first derivatives of the log likelihood and the Hessian matrix is the second derivatives.

The first derivatives with respect to β are given by

$$\frac{\partial}{\partial \beta_j} InL(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i \left(\frac{\boldsymbol{x_{ij}}}{1 + e^{\mathbf{x_i \beta}}} \right) + (1 - y_i) \left(\frac{-\boldsymbol{x_{ij}} e^{\mathbf{x_i \beta}}}{1 + e^{\mathbf{x_i \beta}}} \right) \right) \tag{8}$$

$$= \sum_{i=1}^{n} \left(y_i \boldsymbol{x_{ij}} \left(\frac{1}{1 + e^{\mathbf{x_i \beta}}} \right) - (1 - y_i) \boldsymbol{x_{ij}} \left(\frac{\mathbf{e^{x_i \beta}}}{1 + \mathbf{e^{x_i \beta}}} \right) \right) \tag{9}$$

$$= \sum_{i=1}^{n} \left(y_i \boldsymbol{x_{ij}} (1 - p_i) - (1 - y_i) \boldsymbol{x_{ij}} \, p_i \right) \tag{10}$$

$$= \sum_{i=1}^{n} \left(y_i \boldsymbol{x_{ij}} - y_i \boldsymbol{x_{ij}} \, p_i - \boldsymbol{x_{ij}} \, p_i + y_i \boldsymbol{x_{ij}} \, p_i \right) \tag{11}$$

$$= \sum_{i=1}^{n} \left(\boldsymbol{x_{ij}} (y_i - p_i) \right) \tag{12}$$

where $j = 0, ...d$ and d is the number of parameters. Each of the partial derivatives is then set to zero. In matrix form, equation (12) is written as

$$\mathbf{X^T (y - p)} = 0 \tag{13}$$

Now, the second derivatives with respect to β are given by

$$\frac{\partial^2}{\partial \beta_j \partial \beta_k} InL(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\frac{-\boldsymbol{x_{ij}} \boldsymbol{x_{ik}} e^{\mathbf{x_i \beta}}}{(1 + e^{\mathbf{x_i \beta}})(1 + e^{\mathbf{x_i \beta}})} \right) \tag{14}$$

$$= \sum_{i=1}^{n} \left(-\boldsymbol{x_{ij}} \boldsymbol{x_{ik}} \left(p_i (1 - p_i) \right) \right) \tag{15}$$

If v_i is defined as $p_i(1 - p_i)$ and $\mathbf{V} = diag(v_1, ..., v_n)$ then the Hessian matrix is written as

$$\mathbf{H(\beta)} = \frac{\partial^2}{\partial^2 \beta} InL(\boldsymbol{\beta}) = -\mathbf{X^T V X} \tag{16}$$

Over-fitting the training data may arise in LR [17], especially when the data are very high dimensional and/or sparse. One of the approaches to reduce over-fitting is through quadratic *regularization*, known also as *ridge regression*, which introduces a penalty for large values of β and to obtain better generalization [19]. The regularized log likelihood is defined as

$$InL(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i In \left(\frac{e^{\mathbf{x_i \beta}}}{1 + e^{\mathbf{x_i \beta}}} \right) + (1 - y_i) In \left(\frac{1}{1 + e^{\mathbf{x_i \beta}}} \right) \right) - \frac{\lambda}{2} \|\beta\|^2 \tag{17}$$

$$= \sum_{i=1}^{n} \left(y_i \left(In \left(\frac{e^{\mathbf{x_i \beta}}}{1 + e^{\mathbf{x_i \beta}}} \right) - In \left(\frac{1}{1 + e^{\mathbf{x_i \beta}}} \right) \right) + In \left(\frac{1}{1 + e^{\mathbf{x_i \beta}}} \right) \right) - \frac{\lambda}{2} \|\beta\|^2 \tag{18}$$

$$= \sum_{i=1}^{n} \left(y_i \left(In\left(e^{\mathbf{x_i\beta}}\right) - In\left(1 + e^{\mathbf{x_i\beta}}\right) - In\,1 + In\left(1 + e^{\mathbf{x_i\beta}}\right)\right) + In\,1 - In\left(1 + e^{\mathbf{x_i\beta}}\right)\right) - \frac{\lambda}{2}\|\beta\|^2 \quad (19)$$

$$= \sum_{i=1}^{n} \left(y_i \left(In\left(e^{\mathbf{x_i\beta}}\right) - 0\right) + 0 - In\left(1 + e^{\mathbf{x_i\beta}}\right)\right) - \frac{\lambda}{2}\|\beta\|^2 \quad (20)$$

$$= \sum_{i=1}^{n} \left(In\left(e^{\mathbf{x_i\beta}}\right)^{y_i} - ln\left(1 + e^{\mathbf{x_i\beta}}\right)\right) - \frac{\lambda}{2}\|\beta\|^2 \quad (21)$$

$$= \sum_{i=1}^{n} In\left(\frac{e^{y_i\mathbf{x_i\beta}}}{1 + e^{x_i\beta}}\right) - \frac{\lambda}{2}\|\beta\|^2 \quad (22)$$

$$= -\sum_{i=1}^{n} In\left(e^{-y_i\mathbf{x_i\beta}}\left(1 + e^{\mathbf{x_i\beta}}\right)\right) - \frac{\lambda}{2}\|\beta\|^2 \quad (23)$$

where the regularization (penalty) term $\frac{\lambda}{2}\|\beta\|^2$ was added to formula (5). The objective is then to find the Maximum Likelihood Estimate (MLE), $\hat{\beta}$, which maximizes the log likelihood. For binary outputs, the loss function or the deviance DEV is the negative log likelihood and is given by the formula

$$DEV\left(\hat{\beta}\right) = -\,2\text{In L}(\beta) \quad (24)$$

4. Iteratively Re-weighted Least Squares (IRLS)

An alternative to numerically maximizing the LR maximum likelihood equations is the *iteratively re-weighted least squares* (IRLS) technique. This technique uses the Newton-Raphson algorithm to solve the LR score equations. Each- iteration finds the *weighted least squares* (WLS) estimates for a given set of weights, which are used to construct a new set of weights [18]. The gradient and the Hessian are obtained by differentiating the regularized likelihood in (22) with respect to β.

$$\frac{\partial}{\partial\beta}\, InL(\boldsymbol{\beta}) = \mathbf{X^T(y - p)} - \lambda\boldsymbol{\beta} = 0 \quad (25)$$

$$\frac{\partial^2}{\partial^2\beta}\, InL(\boldsymbol{\beta}) = -\mathbf{X^T V X} - \lambda\mathbf{I} = 0 \quad (26)$$

where **I** is a $d \times d$ identity matrix. Now that the first and second derivatives are obtained, the Newton-Raphson update formula on the $(c+1)$–*th* iteration is given by

$$\widehat{\boldsymbol{\beta}}^{(\mathbf{c+1})} = \widehat{\boldsymbol{\beta}}^{(\mathbf{c})} + (\mathbf{X^T V X} + \lambda\mathbf{I})^{-\mathbf{1}}\left(\mathbf{X^T(y - p)} - \lambda\hat{\beta}^{(c)}\right) \quad (27)$$

Since $\widehat{\boldsymbol{\beta}}^{(\mathbf{c})} = (\mathbf{X^T V X} + \boldsymbol{\lambda}\mathbf{I})^{-\mathbf{1}}(\mathbf{X^T V X} + \boldsymbol{\lambda}\mathbf{I})\widehat{\boldsymbol{\beta}}^{(\mathbf{c})}$, then (27) can be rewritten as

$$\widehat{\boldsymbol{\beta}}^{(\mathbf{c+1})} = (\mathbf{X^T V X} + \lambda\mathbf{I})^{-\mathbf{1}}\mathbf{X^T}\left(\mathbf{V X}\widehat{\boldsymbol{\beta}}^{(\mathbf{c})} + (\mathbf{y - p})\right) \quad (28)$$

$$= (\mathbf{X^T V X} + \lambda\mathbf{I})^{-\mathbf{1}}\mathbf{X^T V z}^{(\mathbf{c})} \quad (29)$$

where $\mathbf{z}^{(\mathbf{c})} = \mathbf{X}\widehat{\boldsymbol{\beta}}^{(\mathbf{c})} + \mathbf{V}^{-\mathbf{1}}(\mathbf{y - p})$ and is referred to as the adjusted response [4].

If the matrix $(\mathbf{X^T V X} + \lambda\mathbf{I})$ were dense, the iterative computation could become unacceptably slow [20]. This necessitates the need for a "trade off" between convergence speed and accurate Newton direction [21]. The method which provides such a trade-off is known as the truncated Newton's method.

4.1. Truncated Regularized Iteratively Re-weighted Least Squares (TR-IRLS)

The weighted least squares (WLS) sub problem is then

$$(\mathbf{X^T V X} + \lambda \mathbf{I})\widehat{\boldsymbol{\beta}}^{(c+1)} = \mathbf{X^T V z}^{(c)} \tag{30}$$

is a linear system of *d* equations and variables, and solving it is equivalent to minimizing the quadratic function

$$\frac{1}{2}\widehat{\boldsymbol{\beta}}^{(c+1)}(\mathbf{X^T V X} + \lambda \mathbf{I})\widehat{\boldsymbol{\beta}}^{(c+1)} - \widehat{\boldsymbol{\beta}}^{(c+1)}\mathbf{X^T V z}^{(c)} \tag{31}$$

Komarek and Moore [22] used *truncated-regularized iteratively-reweighted least squares* (TR-IRLS) technique that implemented a modified linear Conjugate Gradient (CG) to approximate the Newton direction in solving the IRLS for LR.

Conjugate gradient (CG) is an iterative minimization algorithm. The main advantage of the CG method is that it guarantees convergence in at most *d* steps [21]. The TR-IRLS algorithm consists of two loops. Algorithm 1 represents the outer loop finds the solution to the WLS problem and is terminated when the relative difference of deviance between two consecutive iterations is no larger than a specified threshold ε_1. Algorithm 2 represents the inner loop, which solves the WLS sub problems in Algorithm 1 through the linear CG method, which is the Newton direction. Algorithm 2 is terminated when the residual

$$\mathrm{r}^{(c+1)} = (\mathbf{X^T V X} + \lambda \mathbf{I})\widehat{\boldsymbol{\beta}}^{(c+1)} - \mathbf{X^T V z}^{(c)} \tag{32}$$

is no greater than a specified threshold ε_2 [6].

5. Logistic Regression in Rare Events Data

King and Zeng [23] recommend two methods of estimation for choice-based sampling, *prior correction* and *weighting*. This study uses weighting method.

Under pure endogenous sampling, the conditioning is on **X** rather than **y** [24, 25], and the joint distribution of **y** and **X** in the sample is

$$f_s(\boldsymbol{y}, \mathbf{X}|\boldsymbol{\beta}) = P_s(\boldsymbol{X}|\boldsymbol{y}, \boldsymbol{\beta})P_s(\boldsymbol{y}) \tag{33}$$

where β is the unknown parameter to be estimated. Yet, since **X** is a matrix of exogenous variables, then the conditional probability of **X** in the sample is equal to that in the population, or $P_s(\boldsymbol{X}|\boldsymbol{y}, \beta) = P(\boldsymbol{X}|\boldsymbol{y}, \beta)$. However, the conditional probability in the population is

$$P(\boldsymbol{X}|\boldsymbol{y}, \boldsymbol{\beta}) = \frac{f(\boldsymbol{y}, \mathbf{X}|\boldsymbol{\beta})}{P(\boldsymbol{y})} \tag{34}$$

but

$$f(\boldsymbol{y}, \mathbf{X}|\boldsymbol{\beta}) = P(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta})P(\boldsymbol{X}) \tag{35}$$

and hence, substituting and rearranging yields

$$f_s(\boldsymbol{y}, \mathbf{X}|\boldsymbol{\beta}) = \frac{P_s(\boldsymbol{y})}{P(\boldsymbol{y})}P(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta})P(\boldsymbol{X}) \tag{36}$$

$$= \frac{H}{Q}P(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta})P(\boldsymbol{X}) \tag{37}$$

where $\frac{H}{Q} = \frac{P_s(\boldsymbol{y})}{P(\boldsymbol{y})}$, with H representing the proportions in the sample and Q the proportions in the population. The likelihood is then

$$L_{Endogenous} = \prod_{i=1}^{n} \frac{H_i}{Q_i} P(\boldsymbol{y_i}|\boldsymbol{x_i}, \boldsymbol{\beta}) P(\boldsymbol{x_i}) \tag{38}$$

where $\frac{H_i}{Q_i} = \left(\frac{\bar{y}}{\tau}\right) y_i + \left(\frac{1-\bar{y}}{1-\bar{\tau}}\right)(1 - y_i)$, with $\bar{y}$ is the proportion of events in the sample and τ is the proportion of events in the population. When dealing with REs and imbalanced data, the likelihood needs to be maximized.

The log-likelihood for LR can then be rewritten as

$$In\, L(\beta|y, \mathbf{X}) = \sum_{i=1}^{n} \frac{Q_i}{H_i} In\, P(y_i|x_i, \beta) \tag{39}$$

$$= \sum_{i=1}^{n} \frac{Q_i}{H_i} In \left(\frac{e^{y_i \mathbf{x_i} \boldsymbol{\beta}}}{1 + e^{x_i \beta}}\right) \tag{40}$$

$$= \sum_{i=1}^{n} w_i In \left(\frac{e^{y_i \mathbf{x_i} \boldsymbol{\beta}}}{1 + e^{\mathbf{x_i} \boldsymbol{\beta}}}\right) \tag{41}$$

where $w_i = \frac{Q_i}{H_i}$. If the proportion of events in the sample is more than that in the population, then the ratio $\frac{Q}{H}$ is less than one, and hence the events are given less weight, while the non-events would be given more weight if their proportion in the sample is less than that in the population. The above estimator, however, is not fully efficient, because the information matrix equality does not hold. This is demonstrated as

$$-\mathrm{E}\left[\frac{Q}{H} \nabla_{\beta}^2 P(\boldsymbol{y}|\boldsymbol{X}, \beta)\right] \neq \mathrm{E}\left[\left(\frac{Q}{H} \nabla_{\beta} P(\boldsymbol{y}|\boldsymbol{X}, \beta)\right)\left(\frac{Q}{H} \nabla_{\beta} P(\boldsymbol{y}|\boldsymbol{X}, \beta)\right)^{T}\right] \tag{42}$$

and for the LR model it is

$$-\left[\frac{1}{n} \sum_{i=1}^{n} \left(\frac{Q_i}{H_i}\right) p_i(1 - p_i) \boldsymbol{x_i} \boldsymbol{x_j}\right] \neq \left[\frac{1}{n} \sum_{i=1}^{n} \left(\frac{Q_i}{H_i}\right)^2 p_i(1 - p_i) \boldsymbol{x_i} \boldsymbol{x_j}\right] \tag{43}$$

Let $\boldsymbol{A} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{Q_i}{H_i}\right) p_i(1 - p_i) \boldsymbol{x_i} \boldsymbol{x_j}$, and $\boldsymbol{B} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{Q_i}{H_i}\right)^2 p_i(1 - p_i) \boldsymbol{x_i} \boldsymbol{x_j}$, then the asymptotic variance matrix of the estimator β is given by the *sandwich estimate*, such that $\boldsymbol{V}(\beta) = \boldsymbol{A^{-1} B A^{-1}}$ [26, 27, 28].

King and Zeng [23] extended the small-sample bias corrections, as described by McCullagh and Nelder [29] to include the weighted likelihood (41). According to McCullagh and Nelder [29], and later Cordeiro and McCullagh [30], the bias vector is given by

$$\mathrm{bias}(\hat{\beta}) = (\mathbf{X^T V X})^{-1} \mathbf{X^T V} \xi \tag{44}$$

where and Q_{ii} are the diagonal elements of $\mathbf{Q} = \mathbf{X}(\mathbf{X^T D X})^{-1} \mathbf{X^T}$ which is the approximate covariance matrix of the logistic link function. The second-order bias-corrected estimator is then

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \mathbf{B}(\hat{\boldsymbol{\beta}}) \tag{45}$$

As for the variance matrix $\mathbf{V}(\tilde{\beta})$ of $\tilde{\beta}$, it is estimated using

$$\mathbf{V}(\tilde{\boldsymbol{\beta}}) = \left(\frac{n}{n + d}\right)^2 \mathbf{V}(\hat{\boldsymbol{\beta}}) \tag{46}$$

Since $\left(\frac{n}{n+d}\right)^2 < 1$, then $\mathbf{V}(\tilde{\boldsymbol{\beta}}) < \mathbf{V}(\hat{\boldsymbol{\beta}})$, and hence both the variance and the bias are now reduced.

6. Rare Event Weighted Logistic Regression (RE-WLR)

Now the formulation of the weighted LR suggested by King and Zeng [23] is applied to the WLR model in (41), the weighted likelihood can be rewritten as

$$L_w(\beta) = \prod_{i=1}^{n} (p_i)^{w_1 y_i} (1-p_i)^{w_0 (1-y_i)} \tag{47}$$

where $w_1 = \frac{\tau}{\bar{y}}$. $w_0 = \frac{1-\tau}{1-\bar{y}}$, $p_i = \frac{1}{1+e^{-\eta_i}}$, and $(1-p_i) = \frac{e^{-\eta_i}}{1+e^{-\eta_i}}$.

Now,

$$p_i = \left(\frac{1}{1+e^{-\eta_i}}\right)^{w_1} = {p_i}^{w_1} \tag{48}$$

and hence,

$$p_i{}' = \frac{\partial p_i}{\partial \eta_i} = w_1 \left(\frac{1}{1+e^{-\eta_i}}\right)^{w_1 - 1} (-1)\left(1+e^{-\eta_i}\right)^{-2}\left(-e^{-\eta_i}\right) \tag{49}$$

$$= w_1 \left(\frac{1}{1+e^{-\eta_i}}\right)^{w_1} \left(1+e^{-\eta_i}\right)\left(1+e^{-\eta_i}\right)^{-2}\left(e^{-\eta_i}\right) \tag{50}$$

$$= w_1 \left(\frac{1}{1+e^{-\eta_i}}\right)^{w_1} \left(\frac{e^{-\eta_i}}{1+e^{-\eta_i}}\right) \tag{51}$$

$$= w_1 \, {p_i}^{w_1} (1-p_i) \tag{52}$$

$$p_i{}'' = \frac{\partial^2 p_i}{\partial {\eta_i}^2} = w_1^2 \left(\frac{1}{1+e^{-\eta_i}}\right)^{w_1} \left(\frac{e^{-\eta_i}}{1+e^{-\eta_i}}\right)^2 + w_1 \left(\frac{1}{1+e^{-\eta_i}}\right)^{w_1} \left(\frac{-e^{-\eta_i}}{1+e^{-\eta_i}} + \frac{e^{-\eta_i}(-e^{-\eta_i})}{(1+e^{-\eta_i})^2}\right) \tag{53}$$

$$= w_1^2 \left(\frac{1}{1+e^{-\eta_i}}\right)^{w_1} \left(\frac{e^{-\eta_i}}{1+e^{-\eta_i}}\right)^2 + w_1 \left(\frac{1}{1+e^{-\eta_i}}\right)^{w_1} \left(\frac{e^{-\eta_i}(-1-e^{-\eta_i}+e^{-\eta_i})}{(1+e^{-\eta_i})^2}\right) \tag{54}$$

$$= w_1^2 \left(\frac{1}{1+e^{-\eta_i}}\right)^{w_1} \left(\frac{e^{-\eta_i}}{1+e^{-\eta_i}}\right)^2 + w_1 \left(\frac{1}{1+e^{-\eta_i}}\right)^{w_1} \left(\left(\frac{e^{-\eta_i}}{1+e^{-\eta_i}}\right)\left(-\frac{1}{1+e^{-\eta_i}}\right)\right) \tag{55}$$

$$= w_1 \, {p_i}^{w_1} (1-p_i) w_1 (1-p_i) + w_1 \, {p_i}^{w_1} (1-p_i)(-p_i) \tag{56}$$

$$= w_1 \, {p_i}^{w_1} (1-p_i)(w_1 - w_1 \, p_i) + w_1 \, {p_i}^{w_1} (1-p_i)(-p_i) \tag{57}$$

$$= w_1 \, {p_i}^{w_1} (1-p_i)(w_1 - (1+w_1) p_i) \tag{58}$$

Finally, the bias vector for WLR can now be rewritten as

$$\mathbf{B}(\hat{\boldsymbol{\beta}}) = (\mathbf{X^T D X})^{-1} \mathbf{X^T D} \xi \tag{59}$$

where the i-th element of the vector $\boldsymbol{\xi}$ is now

$$\xi_i = 0{,}5 Q_{ii} \left((1+w_1) p_i - w_1\right) \tag{60}$$

with Q_{ii} as the diagonal elements of Q, which is now $\mathbf{Q} = \mathbf{X}(\boldsymbol{X^T D X})^{-1} \boldsymbol{X^T}$, and $\mathbf{D} = diag\,(v_i \, w_i)$ for $i = 1 \ldots n$. The bias-corrected estimator becomes

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \mathbf{B}(\hat{\boldsymbol{\beta}}) \tag{61}$$

Iteratively re-weighted least squares (IRLS) method is used to find the MLE of β, which uses Newton-Raphson algorithm to solve LR score equations. Each iteration finds the *weighted least squares* (WLS)

estimates for a given set of weights, which are used to construct a new set of weights. For WLR, the gradient and Hessian are obtained by differentiating the regularized weighted log-likelihood,

$$In\, L_w(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i \; In\; \frac{e^{y_i \mathbf{x_i}\boldsymbol{\beta}}}{1+e^{\mathbf{x_i}\boldsymbol{\beta}}} - \frac{\lambda}{2}\|\beta\|^2 \tag{62}$$

with respect to β. In matrix form, the gradient is

$$\frac{\partial}{\partial \beta_j}\, InL_w(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i \left(\left(\frac{1+e^{\mathbf{x_i}\boldsymbol{\beta}}}{e^{y_i x_i \beta}} \right) \left(\left(\frac{y_i \boldsymbol{x_{ij}} e^{y_i \mathbf{x_i}\boldsymbol{\beta}}}{1+e^{\mathbf{x_i}\boldsymbol{\beta}}} \right) + \left(\frac{e^{y_i \mathbf{x_i}\boldsymbol{\beta}} \left(-\boldsymbol{x_{ij}} e^{\mathbf{x_i}\boldsymbol{\beta}}\right)}{(1+e^{\mathbf{x_i}\boldsymbol{\beta}})^2} \right) \right) \right) - \lambda\beta \tag{63}$$

$$= \sum_{i=1}^{n} w_i \left(\left(\frac{1+e^{\mathbf{x_i}\boldsymbol{\beta}}}{e^{y_i x_i \beta}} \right) \left(\left(\frac{y_i \boldsymbol{x_{ij}} e^{y_i \mathbf{x_i}\boldsymbol{\beta}} \left(1+e^{\mathbf{x_i}\boldsymbol{\beta}}\right)}{(1+e^{\mathbf{x_i}\boldsymbol{\beta}})^2} \right) + \left(\frac{e^{y_i \mathbf{x_i}\boldsymbol{\beta}} \left(-\boldsymbol{x_{ij}} e^{x_i \beta}\right)}{(1+e^{\mathbf{x_i}\boldsymbol{\beta}})^2} \right) \right) \right) - \lambda\beta \tag{64}$$

$$= \sum_{i=1}^{n} w_i \left(\left(\frac{1+e^{\mathbf{x_i}\boldsymbol{\beta}}}{e^{y_i \mathbf{x_i}\boldsymbol{\beta}}} \right) \left(\frac{\boldsymbol{x_{ij}} e^{y_i \mathbf{x_i}\boldsymbol{\beta}} \left(y_i\left(1+e^{\mathbf{x_i}\boldsymbol{\beta}}\right) - e^{\mathbf{x_i}\boldsymbol{\beta}}\right)}{(1+e^{\mathbf{x_i}\boldsymbol{\beta}})^2} \right) \right) - \lambda\beta \tag{65}$$

$$= \sum_{i=1}^{n} w_i \left(\boldsymbol{x_{ij}} \left(\frac{\left(y_i\left(1+e^{\mathbf{x_i}\boldsymbol{\beta}}\right) - e^{\mathbf{x_i}\boldsymbol{\beta}}\right)}{1+e^{\mathbf{x_i}\boldsymbol{\beta}}} \right) \right) - \lambda\beta \tag{66}$$

$$= \sum_{i=1}^{n} w_i \boldsymbol{x_{ij}} (y_i - p_i) - \lambda\beta \tag{67}$$

$$\nabla_\beta In\, L_w(\beta) = \boldsymbol{X^T W}(\boldsymbol{y}-\boldsymbol{p}) - \lambda\boldsymbol{\beta} \tag{68}$$

where $\boldsymbol{W} = diag\,(w_i)$ and $\mathbf{p}$ is the probability vector. The Hessian with respect to β is then

$$\frac{\partial^2}{\partial \beta_j \partial \beta_k}\, InL_w(\beta) = \sum_{i=1}^{n} -w_i \boldsymbol{x_{ij}} \left(\left(\frac{\boldsymbol{x_{ik}} e^{\mathbf{x_i}\boldsymbol{\beta}}}{1+e^{x_i\beta}} \right) + \left(\frac{\boldsymbol{x_{ik}} e^{\mathbf{x_i}\boldsymbol{\beta}} \left(-e^{\mathbf{x_i}\boldsymbol{\beta}}\right)}{(1+e^{x_i\beta})^2} \right) \right) - \lambda\mathbf{I} \tag{69}$$

$$= \sum_{i=1}^{n} -w_i \boldsymbol{x_{ij}} \left(\frac{\boldsymbol{x_{ik}} e^{\mathbf{x_i}\boldsymbol{\beta}} \left(1+e^{\mathbf{x_i}\boldsymbol{\beta}} - e^{\mathbf{x_i}\boldsymbol{\beta}}\right)}{(1+e^{\mathbf{x_i}\boldsymbol{\beta}})(1+e^{\mathbf{x_i}\boldsymbol{\beta}})} \right) - \lambda\mathbf{I} \tag{70}$$

$$= \sum_{i=1}^{n} \left(-\boldsymbol{x_{ij}} \boldsymbol{x_{ik}}\, w_i\left(\, p_i(1-p_i)\right) \right) - \lambda\mathbf{I} \tag{71}$$

$$\nabla_\beta^2 In\, L_w(\beta) = -\boldsymbol{X^T D X} - \lambda\mathbf{I} \tag{72}$$

The first and second derivatives are obtained, the Newton-Raphson methods iterates via

$$\widehat{\boldsymbol{\beta}}^{(\mathbf{c+1})} = \widehat{\boldsymbol{\beta}}^{(\mathbf{c})} - (\nabla_\beta^2 In\, L_w(\beta))^{-\mathbf{1}} \left(\nabla_\beta In\, L_w(\beta) \right) \tag{73}$$

$$\widehat{\boldsymbol{\beta}}^{(\mathbf{c+1})} = \widehat{\boldsymbol{\beta}}^{(\mathbf{c})} + (\mathbf{X^T D X} + \lambda\mathbf{I})^{-\mathbf{1}} \left(\mathbf{X^T W}(\mathbf{y}-\mathbf{p}) - \lambda\widehat{\boldsymbol{\beta}}^{(\mathbf{c})} \right) \tag{74}$$

Since $\widehat{\boldsymbol{\beta}}^{(\mathbf{c})} = (\mathbf{X^T D X} + \lambda\mathbf{I})^{-\mathbf{1}} (\mathbf{X^T D X} + \lambda\mathbf{I})\widehat{\boldsymbol{\beta}}^{(\mathbf{c})}$, then (74) can be rewritten as

$$\widehat{\boldsymbol{\beta}}^{(\mathbf{c+1})} = (\mathbf{X^T D X} + \lambda\mathbf{I})^{-\mathbf{1}} (\mathbf{X^T D X} + \lambda\mathbf{I})\widehat{\boldsymbol{\beta}}^{(\mathbf{c})} + (\mathbf{X^T D X} + \lambda\mathbf{I})^{-\mathbf{1}} \left(\mathbf{X^T W}(\mathbf{y}-\mathbf{p}) - \lambda\widehat{\boldsymbol{\beta}}^{(\mathbf{c})} \right) \tag{75}$$

$$\widehat{\boldsymbol{\beta}}^{(\mathbf{c+1})} = (\mathbf{X^T D X} + \lambda\mathbf{I})^{-\mathbf{1}} \mathbf{X^T D} \left(\mathbf{X}\widehat{\boldsymbol{\beta}}^{(\mathbf{c})} + (\mathbf{X^T})^{-\mathbf{1}} \mathbf{D^{-1}} \lambda\widehat{\boldsymbol{\beta}}^{(\mathbf{c})} + \mathbf{D^{-1}} \mathbf{W}(\mathbf{y}-\mathbf{p}) - (\mathbf{X^T})^{-\mathbf{1}} \mathbf{D^{-1}} \lambda\widehat{\boldsymbol{\beta}}^{(\mathbf{c})} \right) \tag{76}$$

$$\widehat{\boldsymbol{\beta}}^{(\mathbf{c+1})} = (\mathbf{X^T D X} + \lambda\mathbf{I})^{-\mathbf{1}} \mathbf{X^T D} \left(\mathbf{X}\widehat{\boldsymbol{\beta}}^{(\mathbf{c})} + \mathbf{D^{-1}} \mathbf{W}(\mathbf{y}-\mathbf{p}) \right) \tag{77}$$

The Newton–Raphson update with respect to β on the $(c+1)$ th iteration is

$$\hat{\beta}^{(c+1)} = (\mathbf{X^T D X} + \lambda \mathbf{I})^{-1} \mathbf{X^T D z^{(c)}} \quad (78)$$

where $\mathbf{z^{(c)}} = \mathbf{X}\hat{\boldsymbol{\beta}}^{(c)} + \mathbf{D^{-1}W(y - p)}$ is the adjusted dependent variable or the adjusted response.

The weighted least squares (WLS) sub problem is then

$$(\mathbf{X^T D X} + \lambda \mathbf{I})\hat{\boldsymbol{\beta}}^{(c+1)} = \mathbf{X^T D z^{(c)}} \quad (79)$$

which is a system of linear equations with a matrix **X**, a vector of adjusted responses **z**, and a weight matrix **D**. Both the weights and the adjusted response vector are dependent on $\widehat{\boldsymbol{\beta}}^{(c)}$, which is the current estimate of the parameter vector. Specifying an initial estimate $\widehat{\boldsymbol{\beta}}^{(0)}$ for $\widehat{\boldsymbol{\beta}}$ can be solved iteratively, giving a sequence of estimates that converges to the MLE of $\widehat{\boldsymbol{\beta}}$. This iterative process can be done using the CG method. Like the TR-IRLS algorithm, in order to avoid the long computations that the CG may suffer from, a limit can be placed on the number of CG iterations, thus creating an approximate or truncated Newton direction [6].

7. Conclusion

Rare Event Weighted Logistic Regression (RE-WLR**)** is a combination of LR rare events correction [23] and TR-IRLS [22]. Rare event correction in LR is applied by setting weighting to LR with the result that Weighted Logistic Regression (WLR) will be formed. Regularization was added to reduce over-fitting. The estimation of β was performed by using the method of *maximum likelihood* (ML), but there was no closed form solution to maximize $InL(\beta)$ with respect to β. Therefore WLR *maximum likelihood estimates* (MLE) was obtained by using IRLS method of Newton-Raphson algorithm to solve WLR score equations. Each of iterations found the *weighted least squares* (WLS) estimating to give set of weights, which were used to construct a new set of weights. The gradient and the Hessian were obtained by differentiating the regularized weighted likelihood with respect to β. In order to solve large optimization problems, Truncated-Newton method was applied.

References

[1] J. Han and M. Kamber. *Data Mining Concepts and Techniques.* Academic Press, 2001.

[2] J. Pacheco, S. Casado, and L. Núñez. A variable selection method based on Tabu search for logistic regression models. *European Journal of Operational Research,* 199:506–511, 2009.

[3] G.-X. Yuan, C.-H. Ho, C.-J. Lin. Recent advances of large-scale linear classification. *Proc. IEEE 100* (9), page 2584–2603, 2012.

[4] T. Hastie, R. Tibshirani and J. Friedman. *The Elements of Statistical Learning*. Springer Verlag, 2001.

[5] P. Karsmakers, K. Pelckmans and J.A.K. Suykens. Multi-class kernel logistic regression: a fixed-size implementation. *Int. Joint Conf. Neural Network*, page 1756–1761, 2007.

[6] M. Maalouf and M. Siddiqi. Weighted logistic regression for large-scale imbalanced and rare events data. *Journal of Knowledge-Based Systems*, 59:141–148, 2014.

[7] D. Lewis and J. Carlett. Heterogeneous Uncertainly Sampling for Supervised Learning. *Proceedings of ICML-94, 11th International Conference on Machine Learning,* Eds: Cohen, W. dan Hirsh, H., Morgan Kaufmann, San Fransisco, page 148–156, 1994.

[8] M. Kubat, S. Matwin, and R. Holte. Machine Learning for the Detection of Oil Spills In Satellite Radar Images. *Machine Learning,* 30:195-215, 1998.

[9] I. Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine,* 23:89–109, 2001.

[10] Chawla, et.al. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence and Research,* 16:321-357, 2002.

[11] G. Wu and E. Chang. Class-Boundary Alignment for Imbalanced Dataset Learning. *ICML 2003 Workshop on Learning from Imbalanced Data Sets II*, Washington, 2003.

[12] C.S. Hilas. Designing Expert System for Fraud Detection in Private Telecommunication Networks. *Expert Systems with Applications,* 36 (9):11559-11569, 2009.

[13] R. Longadge, S. Dongre, and L. Malik. Class Imbalance Problem in Data Mining: Review. *International Journal of Computer Science and Network,* Vol. 2, 2013.

[14] Maalouf and Trafalis. Robust weighted kernel logistic regression in imbalanced and rare events data. *Computational Statistics and Data Analysis 2011*, 55:168–183, 2010.

[15] S.P. Rahayu. *Logistic regression methods for classification of imbalanced data*. Tesis Ph.D, University Malaysia Pahang (UMP), Pahang, 2012.

[16] Wang, Xu, and Zhou. Large Unbalanced Credit Scoring Using Lasso-Logistic Regression Ensemble. *PLoS ONE 10(2): e0117844*, 2015.

[17] D.W. Hosmer and S. Lemeshow. *Applied Logistic Regression*, *second ed*.. Wiley, 2000.

[18] P. Garthwaite, I. Jolliffe, and J. Byron, J. *Statistical Inference*. Oxford University Press, 2002.

[19] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[20] P. Komarek. *Logistic regression for data mining and high-dimensional classification*. Ph.D. thesis, Carnegie Mellon University, 2004.

[21] J.M. Lewis, S. Lakshmivarahan, S. Dhall. *Dynamic Data Assimilation: A Least Squares Approach*. Cambridge University Press, 2006.

[22] P. Komarek and A. Moore, A. *Making logistic regression a core data mining tool: A practical investigation of accuracy, speed, and simplicity*. Tech. rep., Carnegie Mellon University, 2005.

[23] G. King and L. Zeng. Logistic regression in rare events data. *Political Analysis,* 9:137–163, 2001.

[24] A.C. Cameron, and P.K. Trivedi. *Microeconometrics: Methods and Applications*. Cambridge University Press, 2005.

[25] M. Milgate, J. Eatwell, and P.K Newma. *Econometrics*. W. W. Norton & Company (1990)

[26] T. Amemiya. *Advanced Econometrics*. Harvard University Press, 1985.

[27] Y. Xie and C.F. Manski. The logit model and response-based samples. *Sociological Methods & Research*, 17**:**283–302, 1989.

[28] C.F. Manski and S.R. Lerman. The estimation of choice probabilities from choice based samples. *Econometrica,* 45:1977–1988 ,1977,

[29] P. McCullagh and J. Nelder. *Generalized Linear Model*. Chapman and Hall/CRC, 1989.

[30] G.M. Cordeiro and P. McCullagh. Bias correction in generalized linear models, *Journal of Royal Statistical Society*, 53(3):629–643, 1991.