The use of entropy based fuzzy membership on weighted logistic regression for the unbalanced data

Ajiwasesa Harumeka\textsuperscript{1,2*}, Santi Wulan Purnami\textsuperscript{1}, Santi Puteri Rahayu\textsuperscript{1}

\textsuperscript{1}Department of Statistics, Faculty of Science and Data Analytics, Sepuluh Nopember Institute of Technology, Surabaya, Indonesia
\textsuperscript{2}Statistics Indonesia, Jakarta, Indonesia

*Corresponding E-mail : aharumeka@gmail.com; ajiwasesa@bps.go.id

**Abstract.** Logistic regression is a popular and powerful classification method. The addition of ridge regularization and optimization using a combination of linear conjugate gradients and IRLS, called Truncated Regularized Iteratively Re-weighted Least Square (TR-IRLS), can outperform Support Vector Machine (SVM) in terms of processing speed, especially when applied to large data and have competitive accuracy. However, neither SVM nor TR-IRLS is good enough when applied to unbalanced data. Fuzzy Support Vector Machine (FSVM) is an SVM development for unbalanced data that adds fuzzy membership to each observation. The fuzzy membership makes the interest of each observation in the minority class higher than the majority class. Meanwhile, TR-IRLS developed into a Rare Event Weighted Logistic Regression (RE-WLR) by adding weight to logistic regression and bias correction. The weighting of the RE-WLR depends on the undersampling scheme. It allows an "information loss". Between FSVM and RE-WLR has a similarity, the weight based only on class differences (minority or majority). Entropy Based Fuzzy Support Vector Machine (EF SVM) is a method used to accommodate the weaknesses of FSVM by considering the class certainty of class observations. As a result, EF SVM is able to improve SVM performance for unbalanced data, even beating FSVM. For this reason, we use EF on the TR-IRLS algorithm to classify large and unbalanced data, as a proposed method. This method is called Entropy-Based Fuzzy Weighted Logistic Regression (EF-WLR). This Research shows the review of EF-WLR for unbalanced data classification.

**Keywords:** Classification, Entropy Based Fuzzy, Logistic Regression, Unbalanced Dataset

**Track Name:** Advanced Technology and Renewable Energy

1. **Introduction**

Logistic regression is a popular method. This method, as a linear classifier, has several advantages [1]. First, when compared to kernel-based methods, logistic regression is implemented directly on the data in the original dimension, without being transformed into a higher dimension. Second, logistic regression as a linear classifier proved to be a robust classification method by providing probability values and covering multi-class classification problems. Third, this method does not require...
assumptions about the distribution of the response variables. Fourth, there are no constraints to solve the optimization problem.

Komarek and Moore [2] introduced Truncated Regularized Iteratively Re-weighted Least Square (TR-IRLS) which applies Truncated Newton for large-scale data. TR-IRLS is faster than SVM. Both methods are comparable because they have similar loss functions [3]. Then, Truncated Newtons are used as an optimization method for other logistic regression problems [4]. Both TR-IRLS and SVM are designed with the assumption of balanced data conditions. When in fact, there are incidents of unbalanced data, such as tornado storms, forest covertype [1], collon and breast cancer [5], and State Failure [6].

There are two approaches to overcome the problem of unbalanced data, namely the data level approach and the algorithm level approach [7]. The algorithm-level approach has advantages over the data-level approach. It is an effective method and does not need to change the distribution of training data [7]. Then, SVM and TR-IRLS developed for unbalanced data problems using an algorithm level approach. Maalouf and Siddiqi [1], in 2014, developed the TR-IRLS to overcome unbalanced data, named Rare Event Weighted Logistic Regression (RE-WLR), by combining TR-IRLS, weighting, and bias correction [8]. RE-WLR has better accuracy than TR-IRLS. While, Lin and Wang [9], in 2002, added an element of fuzzy membership to each observation as weight and reformulated SVM, named Fuzzy Support Vector Machine (FSVM). FSVM can outperform SVM in unbalanced data. Both RE-WLR and FSVM have similar weights. The weighting value of the two methods is based on class differences only. So that, all observations in the majority class only has the same value.

Fan, et.al [10] propose Entropy Based Fuzzy Support Vector Machine (EFSVM) to handle weaknesses in FSVM by considering the class certainty of each observation. The certainty of the class is measured using entropy. The smaller the entropy value, the greater the class certainty [10]. The entropy values on the majority class observations will be grouped into subsets. Then we get the fuzzy membership value in each subset. They show that weighting using entropy based fuzzy membership (EF) can improve SVM performance when applied to unbalanced data, even beating FSVM.

Based on the success of entropy based fuzzy membership (EF) on SVM, we propose a method that applies EF to TR-IRLS algorithm, named Entropy Based Fuzzy Weighted Logistic Regression (EF-WLR). By implementing EF, the TR-IRLS algorithm is expected to be able to pay more attention to observations that have a higher class certainty value and can be implemented on large scale data.

2. Regularized Logistic Regression

Logistic regression is a regression analysis method in which the response variable is a categorical variable [11]. If the response variable has only two categories, then it called binary logistic regression. Let there are \( n \) mutually exclusive samples. From each sample there are \( \overrightarrow{X}_i \) and \( y_i \) pairs, \( i = 1, 2, \ldots, n \). \( \overrightarrow{X}_i \) is a vector of size \( p \) which is the number of predictor variables. \( y_i \) is the value of the response variable in the form of categorical 0 or 1. Let sample with response \( y_i = 0 \) belong to a negative class or majority class and sample with response \( y_i = 1 \) belong to a positive class or minority class. Then, \( y_i \) has a Bernoulli distribution with a probability function

\[
P(y_i/n_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i},
\]

\( \pi_i \) is the probability of the \( i \)-th occurrence. The logistic regression equation is

\[
ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \overrightarrow{X}_i \cdot \overline{\beta}
\]

\( \overline{\beta} \) is a parameter vector of size \( p \). Then \( \pi_i \) will become
The parameter estimation method that is often used is Maximum Likelihood (ML). For that, the ln-likelihood function of the variable with Bernoulli distribution is

$$\ln(L(\beta_*)) = \sum_{i=1}^{n} \ln(p_i, \pi_i(1-p_i)^{1-\pi_i}) \tag{4}$$

by substituting equation (3) into equation (4), then we get ln-likelihood function of logistic regression as follows

$$\ln(L(\beta_*)) = \sum_{i=1}^{n} \left( y_i \ln \left( \frac{e^{x_i^T \beta_*}}{1 + e^{x_i^T \beta_*}} \right) + (1 - y_i) \ln \left( \frac{1}{1 + e^{x_i^T \beta_*}} \right) \right) \tag{5}$$

There is an overfitting problem in modeling, the model performs well on training data but poorly on testing data. Adding regularization is one way to reduce these problems [12]. The regularization process limits the influence of the predictor variables that do not affect the response variables. Adding regularization on ln-likelihood function produce regularized ln-likelihood as follows

$$\ln \left( L(\beta) \right) = \sum_{i=1}^{n} \left( \ln \left( \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} \right) \right) - \frac{\lambda}{2} \| \beta \|^2 \tag{6}$$

where $\| \beta \|^2 = \sum_{j=1}^{p} \beta_j^2$ and $\lambda$ is regularization parameter. The regression model in equation (6) usually called ridge regression. Then, the objective is to maximize the likelihood function on equation (6) to get $\beta$, which is estimated value of $\beta$. For binary logistic regression, there is a statistical measure called Deviance [11]. Deviance has the same role as Sum Square Error (SSE) in linear regression. The value of Deviance is obtained by formula

$$\text{Deviance} = -2 \ln \left( L(\hat{\beta}) \right) \tag{7}$$

Maximizing equation (6) with respect to $\beta$ results in a solution that is not closed form. Therefore, numerical method is used. One of the numeric iteration methods is Iteratively Re-weighted Least Square (IRLS) which uses the Newton-Rhapson method.

The Newton-Rhapson method requires a gradient vector and a hessian matrix which are the first and second derivatives of the likelihood function on equation (6), respectively. The gradient vector is

$$G(\beta) = \frac{d \ln(L(\beta))}{d \beta} = X^T(\bar{y} - \pi) - \lambda \beta \tag{8}$$

While the hessian matrix is

$$H(\beta) = \frac{d^2 \ln(L(\beta))}{d \beta^2} = -X^TXX - \lambda I \tag{9}$$

where $v_i = \pi_i(1-\pi_i)$, $V = \text{diag}(v_1, v_2, ..., v_n)$, and $I$ is identity matrix size $p \times p$. Then the update Newton Rhapson iteration for $\beta$ on $(c+1)$-iteration is
Since then, where is adjusted response.

It becomes problematic if the number of observations is large. The large unit of observation causes the Hessian matrix to be formed to be large too. This requires a long processing time to compute the inverse of the Hessian matrix. Komarek [13], first introduced Truncated Regularized Iteratively Re-weighted Least Square (TR-IRLS) in logistic regression to handle large datasets, using Truncated Newtons with linear Conjugate Gradient (CG) to obtain estimates of . Truncated Newton is a flexible and powerful method for optimizing large data [4][14].

The TR-IRLS method has two iterations. First, find the solution of the Weighted Least Square (WLS) and stop it when the relative difference of deviance statistics of two consecutive iterations does not exceed the specified limit (). The WLS sub problem is

\[
(X^TVX + \lambda I)\hat{\beta}^{(c+1)} = X^TV\hat{Z}^{(c)}
\]  

Whereas the second one looks for a solution to the equation by the CG method [15] and stops the iteration when the residual value, , does not exceed a predetermined limit (). 

3. Weighted Logistic Regression for Large Scale and Unbalanced Data

In order to handle unbalanced data, Maalouf and Siddiqi [1], based on King and Zeng [8], use weighting on logistic regression. If weight is added, the regularized ln-likelihood function becomes

\[
ln\left(L_w(\beta)\right) = \sum_{i=1}^{n} w_i \ln\left(\frac{e^{y_i x_i^T \beta}}{1 + e^{x_i^T \beta}}\right) - \frac{\lambda}{2} \|\beta\|^2
\]  

where \(w_i = \left(\frac{y_i}{\bar{y}}\right) y_i + \left(\frac{1-y_i}{1-\bar{y}}\right)(1 - y_i)\), \(\bar{y}\) is proportion of minority class on population, and \(\bar{y}\) is proportion of minority class on sample. The addition of weight aims to obtain consistent estimators. So that, the minority class is given less weight than the majority [1].

Optimization processes such as those used in TR-IRLS [2] is also used . So that, we need a Gradient vector and a Hessian matrix from the likelihood function in equation (13). The gradient vector become

\[
G(\beta) = \frac{d \ln(L(\beta))}{d \beta} = X^T W (Y - \bar{y}) - \lambda \beta
\]

where \(W = diag(w_1, w_2, ..., w_n)\) and the Hessian Matrix become

\[
H(\beta) = \frac{d \ln(L(\beta))}{d^2 \beta} = -X^T DX - \lambda I
\]

where \(D = diag\{v_1, v_2, v_3, ..., v_n\}\). Then the WLS sub problem is
Regularization in equation (13) will produce bias [1]. So bias correction is required [17]. Maalouf and Siddiqi [1] developed the bias vector described by King and Zeng [8] and Mccullagh and Nelder [18] as follows

\[(X^TDX + \lambda I)\vec{\beta}^{(e+1)} = X^TD\vec{z}^{(e)}\]  

(16)

where \(\vec{\beta}^{(e+1)} = B(\vec{\beta}) = (X^TDX + \lambda I)^{-1}X^TD\vec{z}^{(e)}\)

(17)

where \(\xi_i = 0.5Q_{ii}((1 + w_1)p_i - w_1)\) and \(w_1\) is minority class weight. While \(Q_{ii}\) is the main diagonal of the matrix \(Q\) as follows

\[Q = X(X^TDX + \lambda I)^{-1}X^T\]  

(18)

The bias vector is also optimized using the Truncated IRLS method with linear CG. WLS sub problem for the bias vector is

\[(X^TDX + \lambda I)B(\vec{\beta}) = X^TD\vec{z}^{(e)}\]  

(19)

After obtaining the bias vector, we will get \(\vec{\beta}\) which is the estimation of the bias corrected parameter as follows

\[\vec{\beta} = \vec{\beta} - B(\vec{\beta})\]  

(20)

4. Entropy Based Fuzzy Weighted Logistic Regression (EF-WLR) Algorithm

4.1 Entropy Based Fuzzy Membership

Entropy is an important measure in information theory that can be used to select an event from several events that have a probability [16]. Entropy in a case that has two chances of occurrence is written as

\[H_i = -p_{+i}\ln(p_{+i}) - p_{-i}\ln(p_{-i})\]  

(21)

\(p_{+i}\) and \(p_{-i}\) are the probability of the \(i\)-th training sample \((X_i)\) being the minority class and the majority class, respectively [10]. Both of these probabilities are based on it’s \(k\) nearest neighbors. First, we determine the \(k\) nearest neighbors for each sample. Then, we count the number of minority class (num\(_{+i}\)) and majority class (num\(_{-i}\)) in \(k\) nearest neighbors. The probability of the \(i\)-th training sample being the minority class and the majority class is calculated by

\[p_{+i} = \frac{\text{num}_{+i}}{k}\]  

(22)

\[p_{-i} = \frac{\text{num}_{-i}}{k}\]

Finally, the entropy of \(i\)-th training sample can be calculated by equation (21).

To get fuzzy membership value for majority class [10], first, we select entropy of majority class. Then, separated into \(m\) subset, with increasing order of entropy. Finally, we compute fuzzy membership in each subset by means of

\[FM_i = 1 - \phi(l - 1), l = 1,2,...,m\]  

(23)

where \(FM_i\) is the fuzzy membership value of observation on \(l\)-th subset and \(\phi\) is fuzzy membership parameter which has value \(0 < \phi \leq \frac{1}{m-1}\)

4.2 Using Entropy Based Fuzzy Membership as a Weight on Logistic Regression
In this study, the purpose of weighing the TR-IRLS is to increase the interest of minority class observations, so that the accuracy of the minority class can be increased. It is necessary to give weight to the minority class larger than the majority \[10\]. Then the likelihood function in equation (13) will have a value of \(w_i\) as follow

\[
  w_i = \begin{cases} 
  1, & \text{if } y_i = 1 \\
  FM_i, & \text{if } y_i = 0 \text{ and } X_i \in l-th \text{ subset} 
  \end{cases} \tag{24}
\]

where \(FM_i\) is the fuzzy membership value of observation on \(l\)-th subset which computed by equation (23) and \(0 < FM_i \leq 1\).

The value of \(\overrightarrow{\beta}\) is obtained by Truncated IRLS by Komarek \[2\]. The gradient vector and the hessian matrix have the same form with the equation (14) and (15), respectively. So the WLS sub problem still has the same form as equation (16).

The bias vector \(B(\overrightarrow{\beta})\) still have same form with equation (17). In this study, because of \(w_i\) has value equals to 1, then \(\xi_i\) become

\[
  \xi_i = 0.5Q_{ii}(2\hat{\eta}_i - 1) \tag{25}
\]

Finally, \(\overrightarrow{\beta}\) is calculated using equation (20).

| Algorithm 1. WLR MLE using IRLS |
|----------------------------------|
| **Input**: \(X, Y, \hat{\beta}^{(0)}, B(\overrightarrow{\beta})^{(0)}, m, \varphi, k\) |
| **Output**: \(\overrightarrow{\beta}\) |
| **Begin** |
| \(c = 0\) |
| \(deltadiv = 1\) |
| Get \(\overrightarrow{w}\) using algorithm 2 |
| Do While \(|deltadiv > \varepsilon_1|\) dan \(c \leq maxIRLS\) |
| If \(c > 0\) |
| \(\overrightarrow{\beta}^{(c)} = \overrightarrow{\beta}\) |
| End If |
| For \(i = 1\) to \(n\) |
| \(\hat{\eta}_i = \frac{\exp(X_i^T \overrightarrow{\beta}^{(c)})}{1+\exp(X_i^T \overrightarrow{\beta}^{(c)})}\) |
| \(v_i = \hat{\eta}_i(1 - \hat{\eta}_i)\) |
| \(z_i = \frac{X_i^T \overrightarrow{\beta}^{(c)} + (y_i - \hat{\eta}_i)}{v_i}\) |
| End For |
| \(D = diag(v_i, w_i)\) |
| \(Q = X(X^T DX + \lambda I)^{-1}X^T\) |
| For \(k = 1\) to \(n\) |
| \(\xi_k = 0.5Q_{kk}(2\hat{\eta}_k - 1)\) |
| End For |
| Get \(\overrightarrow{\beta}^{(c+1)}\) using algorithm 3 |
| Get \(B(\overrightarrow{\beta})^{(c+1)}\) using algorithm 4 |
Algorithm 1. WLR MLE using IRLS (continuation)

\[
\tilde{\beta}^{(c+1)} = \tilde{\beta}^{(c+1)} - B \left( \tilde{\beta} \right)^{(c+1)}
\]

\[
deltadev = \frac{\text{deviance}^{(c)} - \text{deviance}^{(c+1)}}{\text{deviance}^{(c+1)}}
\]

\[c = c + 1\]

End While

End

Algorithm 2. Entropy Based Fuzzy Membership to get \( \bar{w} \)

Input : \( X, \bar{Y}, m, \varphi, k \)
Output : \( \bar{w} \)

Begin

For \( i=1 \) to \( n \)

Get \( k \) nearest neighbours of \( \bar{X}_i \)

Count \( num_{+i} \) and \( num_{-i} \) in the \( k \) nearest neighbours

Calculate class probability using equation (22)

\[ H_i = -p_{+i} \ln(p_{+i}) - p_{-i} \ln(p_{-i}) \]

End For

Select entropy of majority class, \( H_\cdot \)

\[ H_{\text{min}} = \min \{ H_\cdot \} \]

\[ H_{\text{max}} = \max \{ H_\cdot \} \]

For \( l=1 \) to \( m \)

\[ \text{thrUp} = H_{\text{min}} + \frac{l}{m} (H_{\text{max}} - H_{\text{min}}) \]

\[ \text{thrLow} = H_{\text{min}} + \frac{l-1}{m} (H_{\text{max}} - H_{\text{min}}) \]

End For

For \( j=1 \) to \( n \)

If \( y_i = \text{majority class} \)

If \( \text{thrLow} \leq H_i < \text{thrUp} \)

\[ FM = 1 - \varphi(l-1) \]

\[ w_i = FM \]

End If

End If

If \( y_i = \text{minority class} \)

\[ w_i = 1 \]

End If

End For

End For

End
Algorithm 3. Linier CG to get $\bar{\beta}$

\begin{align*}
\text{Input} & : A = X^TDX + \lambda I, \quad b = X^TD\bar{z}, \bar{\beta}^{(0)} \\
\text{Output} & : \bar{\beta} \\
\text{Begin} & \\
& t = 0 \\
& \bar{r}^{(0)} = b - A\bar{\beta}^{(0)} \\
& \bar{d}^{(0)} = \bar{r}^{(0)}\\
& \text{Do While } \left\| \bar{r}^{(t)} \right\|^2 > \varepsilon_2 \text{ and } t \leq \text{maxCGBeta} \\
& \quad s^{(t)} = \frac{\bar{r}^{(t)}}{\bar{d}^{(t)T}A\bar{d}^{(t)}} \\
& \quad \bar{\beta}^{(t+1)} = \bar{\beta}^{(t)} + s^{(t)}\bar{d}^{(t)} \\
& \quad \bar{r}^{(t+1)} = \bar{r}^{(t)} - s^{(t)}A\bar{d}^{(t)} \\
& \quad \alpha^{(t)} = \frac{\bar{r}^{(t+1)T}\bar{r}^{(t+1)}}{\bar{r}^{(t)T}\bar{r}^{(t)}} \\
& \quad \bar{d}^{(t+1)} = \bar{r}^{(t+1)} + \alpha^{(t)}\bar{d}^{(t)} \\
& \quad t = t + 1 \\
& \text{End While} \\
\text{End}
\end{align*}

Algorithm 4. Linier CG to get $B(\bar{\beta})$

\begin{align*}
\text{Input} & : A = X^TDX + \lambda I, \quad b = X^TD\xi, B(\bar{\beta})^{(0)} \\
\text{Output} & : B(\bar{\beta}) \\
\text{Begin} & \\
& t = 0 \\
& \bar{r}^{(0)} = b - AB^{(0)} \\
& \bar{d}^{(0)} = \bar{r}^{(0)}\\
& \text{Do While } \left\| \bar{r}^{(t)} \right\|^2 > \varepsilon_2 \text{ and } t \leq \text{maxCGBias} \\
& \quad s^{(t)} = \frac{\bar{r}^{(t)}}{\bar{d}^{(t)T}A\bar{d}^{(t)}} \\
& \quad B(\bar{\beta})^{(t+1)} = B(\bar{\beta})^{(t)} + s^{(t)}\bar{d}^{(t)} \\
& \quad \bar{r}^{(t+1)} = \bar{r}^{(t)} - s^{(t)}A\bar{d}^{(t)} \\
& \quad \alpha^{(t)} = \frac{\bar{r}^{(t+1)T}\bar{r}^{(t+1)}}{\bar{r}^{(t)T}\bar{r}^{(t)}} \\
& \quad \bar{d}^{(t+1)} = \bar{r}^{(t+1)} + \alpha^{(t)}\bar{d}^{(t)} \\
& \quad t = t + 1 \\
& \text{End While} \\
\text{End}
5. Conclusions

Entropy Based Fuzzy Weighted Logistic Regression (EF-WLR) based on Rare Event Weighted Logistic Regression. Weight on weighted logistic regression using entropy based fuzzy membership, allows sample interests not only to be seen based on class differences but also based on class certainty. Regularization process is used to reduce overfitting. Bias correction was applied to obtain unbiased parameters due to the addition of the regularization process. The computation process will be longer if the sample is large. Therefore, truncated IRLS is used to optimize the value of $\hat{\beta}$.

Reference
[1] M Maalouf dan M Siddiqi 2014 “Weighted logistic regression for large-scale imbalanced and rare events data,” Knowledge-Based Syst., vol. 59, hal. 142–148
[2] P Komarek dan A W Moore 2005 “Making logistic regression a core data mining tool with TR-IRLS,” Proc. - IEEE Int. Conf. Data Mining, ICDM, hal. 4–7
[3] J Gareth, D Witten, T Hastie, dan R Tibshirani 2013 An Introduction to Statistical Learning. New York: Springer Science+Business Media
[4] S P Rahayu, J M Zain, A Embong, Juwari, dan S W Purnami 2012 “Logistic regression methods with truncated newton method” Procedia Eng 50 hal. 827–836
[5] F Ladaya, S W Purnami, dan Irhamah 2018 “Entropy Based Fuzzy Support Vector Machine Untuk Klasifikasi Microarray Imbalanced Data” Sepuluh Nopember Institute of Technology, Surabaya
[6] G King dan L Zeng 2001 “Improving Forecasts of State Failure,” World Polit. 53 4. 623–658
[7] A Ali, S M Shamsuddin, dan A L Ralescu 2015 “Classification with class imbalance problem: A review,” Int. J. Adv. Soft Comput. its Appl. 7 3 176–204
[8] G King dan L Zeng 2001 “Logistic regression in rare events data,” J. Stat. Softw. 8 137–163
[9] C F Lin dan S De Wang 2002 “Fuzzy support vector machines,” IEEE Trans. Neural Networks, 13 2 464–471
[10] Q Fan, Z Wang, D Li, D Gao, dan H Zha 2017 “Entropy-based fuzzy support vector machine for imbalanced datasets” Knowledge-Based Syst. 115 87–99
[11] D W Hoshmer dan S Lemeshow 2000 Applied Logistic Regression, Second Edition. John Wiley & Sons , Inc
[12] X Ying 2019 “An Overview of Overfitting and its Solutions,” J. Phys. Conf. Ser. 1168 2
[13] P Komarek 2004 “Logistic Regression for Data Mining and High- Dimensional Classification” Carnegie Mellon University
[14] S G Nash 2000 “A survey of truncated-Newton methods” J. Comput. Appl. Math. 124 45–59
[15] M R Hestenes dan E Stiefel 1934 “Methods of conjugate gradients for solving linear systems,” J. Res. Natl. Bur. Stand. 49 6 409
[16] C E Shannon 1948 “A Mathematical Theory of Communication” Bell Syst. Tech. J. 27 4 623–656