F-measure maximizing logistic regression

Masaaki Okabe (✉), Jun Tsuchida (✉), and Hiroshi Yadohisa (✉)
Graduate School of Culture and Information Science, Doshisha University, Kyoto, Japan

ABSTRACT
Logistic regression is a widely used method in several fields. When applying logistic regression to imbalanced data, wherein the majority classes dominate the minority classes, all class labels are estimated as "majority class." In this study, we use an F-measure optimization method to improve the performance of logistic regression applied to imbalanced data. Although many F-measure optimization methods adopt a ratio of the estimators to approximate the F-measure, the ratio of the estimators tends to exhibit more bias than when the ratio is directly approximated. Therefore, we employ an approximate F-measure to estimate the relative density ratio. In addition, we define and approximate a relative F-measure. We present an algorithm for a logistic regression weighted approximation relative to the F-measure. The results of an experiment using real world data demonstrate that our proposed algorithm can efficiently improve the performance of logistic regression applied to imbalanced data.

1. Introduction
Logistic regression is a binary classification method that consists of a linear model with easily interpretable parameters; therefore, this approach is commonly used in various fields. If the target class imbalance is 1:9 or more, it might be possible to predict all positive results. This problem is caused by the objective function of logistic regression. The objective function yields the error ratio, and we minimize this function when all results are identical.

Imbalanced data are obtained when the ratio of class labels is imbalanced, for example, 1:9. This is observed in various areas, such as credit scoring (Chan and Stolfo 1998; Chang et al. 2016), e-mail filtering (Galar et al. 2012), and medicine (Brown and Mues 2012). In imbalanced data classification, AUC (ROC-AUC) and the F-measures, rather than accuracy, are often used for evaluation (Bradley 1997; Sokolova, Japkowicz, and Szpakowicz 2006). AUC is an evaluation index that expresses the discriminant model classified labels based on order. Therefore, AUC is useful when the threshold is not set. Additionally, AUC is used to evaluate the model overall. By contrast, the F-measure is used to indicate if discriminant models detect minority classes comprehensively and correctly. The F-measure is used as an evaluation index for minority classes to retrieve information (van Rijsbergen 1974). It is often used to categorize class labels as interested classes and other classes, such as loyal customer detection, item purchase, and information retrieval.

In the context of analyzing imbalanced data, other classification methods may predict the majority class; alternatively, logistic regression because evaluation of analysis does not correspond
to logistic regression. Several classification methods for imbalanced data have been proposed by researchers. We can approximately classify these methods into four types. The first are those that use asymmetric misclassification costs (e.g., Elkan 2001; Bahnsen, Aouada, and Ottersten 2014). These methods use an asymmetric loss for misclassification. For example, if the classifier predicts a positive value for a negative case, then the costs will be higher than those of the inverse case. However, defining the optimal cost is different, and misdefining costs can lead to incorrect classification. The second is adjusting class balance by a sampling method (e.g., Chawla et al. 2002; Zhang et al. 2014). This method adjusts class balance by over or under sampling the obtained data. However, a non-linear model is used (e.g., a decision tree) in this method, as a linear model may induce over-fitting. The third is an objective function with a penalizing term (e.g., Firth 1993). This method is often used in logistic regression as a stable objective function. However, because this method does not assume imbalanced data, its accuracy is poor. The final methods are those that use optimizing indices, such as the F-measure (e.g., Musicant, Kumar, and Ozgur 2003; Jansche 2005; Ye et al. 2012) instead of accuracy. These methods optimize a selected index. However, because the objective function is complicated, optimization is not stable.

In this study, we construct a logistic regression that optimizes the F-measure for given imbalanced data and a selected minority class. We adopted the F-measure as we could construct a model for the selected minority class at a certain threshold (= 0.5). In addition, we could optimize the model for the selected minority class and interpret the intercept parameter. Several methods for imbalanced data consider the weight of each sample. For example, the sampling method considers the weight of each sample in the minority class as double that of the majority class. This is the same concept as the cost-sensitive method. In this study, we show it is possible to apply logistic regression to imbalanced data, using the F-measure. Moreover, our method is cost-sensitive for F-measure optimization without domain knowledge, and it does not employ a random number, unlike the sampling method. To optimize the F-measure, we introduce the weight of each sample using the expected F-measure. The sample that contributed to the expected F-measure is heavily weighted. For this sample, we conduct F-measure optimization logistic regression.

This remainder of the paper is structured as follows. After providing some relevant background in Sect. 2, we describe our proposed method in Sect. 3. We then provide a numerical example and real data example in Sect. 4. We conclude this paper and summarize possible future work in Sect. 5.

2. Relevant background

Assume that we are given the dataset \( \{(x_i, y_i)\}_{i=1}^n \), \( x_i \in \mathbb{R}^d \), and \( y_i \in \{0, 1\} \). Let the input vector \( x = (x_1, x_2, ..., x_d)^T \) and class label \( y \) be binary (i.e., \( y \) can be either 1 or 0).

2.1. Logistic regression

Probability \( P(y = 1|x) \) is modeled as follows:

\[
P(y = 1|x) = \frac{1}{1 + \exp(-\beta^T x)},
\]

where \( \beta = (\beta_1, \beta_2, ..., \beta_d)^T \in \mathbb{R}^d \). Assuming \( y_i|x_i \sim B(1, p_i) \), the likelihood function can be written as

\[
\ell(y|\beta) = \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i}.
\]

We can define the error function by taking the negative logarithm of likelihood, which yields a cross-entropy error function \( L(\beta) \) of the following form:
Parameter $b$ is determined such that $L_k$, a regularized cross-entropy error function, is minimized.

$$L_k(b) = \frac{1}{C_0} \sum_{i=1}^{n} \left( y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \right) + \frac{1}{C_0} \log \frac{1}{C_0} p_i,$$

where penalty term $\lambda ||\beta||^2$ is included for regularization purposes and $\lambda (\geq 0)$ denotes the regularization parameter. Minimizing Equation (2) corresponds to maximizing accuracy.

### 2.2 F-measure

In binary classification, an instance is mapped into one of two classes: positive ($y = 1$) or negative ($y = 0$). Table 1 summarizes these outcomes and their notation. The number of positive instances is $n_p = TP + FN$. Similarly, the number of negative instances is $n_n = FP + TN$.

Precision is defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP}.\quad (3)$$

Similarly, recall is defined as

$$\text{Recall} = \frac{TP}{TP + FN}.\quad (4)$$

Precision (Equation (3)) denotes the ratio of the number of true positive (TP) and predicted as positive instances, $m_p$. Similarly, the recall (Equation (4)) denotes ratio TP and the number of positive instances, $n_p$.

The F-measure is defined as

$$F = 2 \left( \frac{1}{\text{Recall}} + \frac{1}{\text{Precision}} \right)^{-1}.\quad (5)$$

This F-measure denotes the harmonic mean of the precision and recall. Similarly, the $\alpha$-relative F-measure is defined as

$$F_\alpha = \left( \alpha \frac{1}{\text{Recall}} + (1 - \alpha) \frac{1}{\text{Precision}} \right)^{-1}$$

$$= \frac{TP}{zn_p + (1 - \alpha)m_p},\quad (6)$$

where $0 \leq \alpha \leq 1$. The relative F-measure denotes the $\alpha$-weighted harmonic mean of the precision and recall. When $\alpha = 1/2$, $F_\alpha$ is equivalent to the F-measure.

**Table 1. Confusion matrix.**

| Predicted Label | 1   | 0   | Total |
|-----------------|-----|-----|-------|
| True Label      |     |     |       |
| 1               | TP  | FN  | $n_p$ |
| 0               | FP  | TN  | $n_n$ |
| Total           | $m_p$ | $m_n$ | $n$   |
Here, we present the method developed for imbalanced data learning. We can approximate the \( a \)-relative F-measure (Equation (6)) by the \( a \)-relative density ratio (Equation (8)). The weight \( w_a(x) \) corresponds to the \( a \)-relative F-measure. When \( a = 1/2 \), the weight \( w_a(x) \) is mapped onto the F-measure.

### 3.1. Underlying idea

The confusion matrix given input vector \( x \) is shown in Table 2. \( p \) shows the probability that the object belongs to positive label \( (y=1) \). \( \hat{p} \) shows the probability that the object is predicted to have the positive label.

From Table 2, the \( a \)-relative F-measure is calculated as follows:

\[
F_a = \frac{q_{TP}}{\alpha p + (1 - \alpha)\hat{p}}.
\]

Considering imbalanced data, the object with a positive label should be predicted to have a positive label. Assuming that \( q_{FN} = 0 \), the \( a \)-relative F-measure is given by the following:

\[
F_a = \frac{p}{\alpha p + (1 - \alpha)\hat{p}}. \quad (7)
\]

Our idea is that Equation (7) is regarded as the weights for the model. However, Equation (7) is not appropriate to use as the weights because directly using \( p \) and \( \hat{p} \) does not consider the frequency of \( p \) and \( \hat{p} \). Considering the frequency of \( p \) and \( \hat{p} \), we decrease the imbalance in label predictions caused by imbalanced class data. Although \( p \) and \( \hat{p} \) are a conditional expectation of \( y_i \) and an estimator of conditional expectation, \( p \) and \( \hat{p} \) may be also regarded as random variables when input vector \( x \) follows a distribution. Accordingly, we adopt a weight for the model as an F-measure represented as a density ratio.

Let weight \( w_a(p) \) be the \( a \)-relative density ratio of \( f_p(p) \) and \( f_{\hat{p}}(p) \):

\[
w_a(p) = \frac{f_p(p)}{\alpha f_p(p) + (1 - \alpha)f_{\hat{p}}(p)} \quad \text{where} \quad f_p(p) = \frac{f_p(p)}{q_a(p)}, \quad (8)
\]

where \( f_p(p) \) and \( f_{\hat{p}}(p) \) are the probability density functions of \( p \) and \( \hat{p} \), respectively. This formulation assumes that \( \{p_i\}_{i=1}^n \) consists of identically distributed (i.i.d.) samples from one-dimensional distribution \( P_p \) with density \( f_p(p) \) and \( \{\hat{p}_i\}_{i=1}^n \) consists of i.i.d. samples from one-dimensional distribution \( P_{\hat{p}} \) with density \( f_{\hat{p}}(p) \):

\[
\{p_i\}_{i=1}^n \overset{i.i.d.}{\sim} P_p, \quad \{\hat{p}_i\}_{i=1}^n \overset{i.i.d.}{\sim} P_{\hat{p}}.
\]

Therefore, we can estimate the weight \( w_a(p) \).
### 3.2. Our method

Our method consists of two steps: a \( \beta \) estimation step and a \( w_x(p) \) estimation step. \( \beta \) is the logistic regression parameter. \( w_x(p) \) is the weight of sample importance corresponding to the F-measure.

#### 3.2.1. Estimating \( \beta \)

When \( \alpha \) is set, we can estimate \( \beta \) with a weighted cross-entropy error function to obtain the following optimization problem:

\[
L_{\beta}(\beta|\theta) = -\sum_{i=1}^{n} \hat{w}_x(p_i) \{ y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \} + \lambda_{\beta}||\beta||^2, \tag{9}
\]

where \( p_i = p(x_i) = 1/(1 + \exp(-\beta^T x_i)) \)

\[
\hat{w}_x(p_i) = \sum_{j=1}^{n} \hat{\theta}_j K(p_i, \hat{p}_j). \]

The weighted cross-entropy error function (Equation (9)) is convex and its analytic solution is the global solution.

Weight \( w_x(p) \) denotes the divergence between the true label and model. Equation (10) shows modeled \( \alpha \)-relative density ratio (Equation (8)). The weight can then be modeled as

\[
w_x(p; \theta) = \sum_{i=1}^{n} \theta_i K(p, \hat{p}_i), \tag{10}
\]

where \( K(p, p_i) \) is a kernel basis function. In this case, we use the Gaussian kernel:

\[
K(p, \hat{p}_i) = \exp \left( \frac{(p - \hat{p}_i)^2}{2\sigma} \right),
\]

where \( \sigma > 0 \) is the kernel width. \( \alpha \) is a tuning parameter of the weight for model, considering non-parametric (i.e., modeling only \( y_i \)) or parametric (i.e., logistic model) options. When \( \alpha = 1 \), because \( w(p) = 1 \), this method is equivalent to a logistic regression. This is the full parametric model. By contrast, when \( \alpha = 0 \), this is a non-parametric model. When \( \alpha = 0.5 \), the F-measure is approximated by \( w(p) \).

Employing a Gaussian kernel, the derivative of the objective function is given by:

\[
\frac{\partial f(y, f)}{\partial \beta_k} = -\sum_{i=1}^{n} \frac{\partial \hat{w}(p(x_i))}{\partial \beta_k} \{ y_i \log(p(x_i)) + (1 - y_i) \log(1 - p(x_i)) \} \\
- \sum_{i=1}^{n} \hat{w}(p(x_i)) \frac{\partial}{\partial \beta_k} \{ y_i \log(p(x_i)) + (1 - y_i) \log(1 - p(x_i)) \} \\
- n \frac{\partial}{\partial \beta_k} \lambda_{\beta} ||\beta||^2. \tag{11}
\]

The first term of Equation (11) is as follows:

\[
- \sum_{i=1}^{n} \frac{\partial \hat{w}(p(x_i))}{\partial \beta_k} \{ y_i \log(p(x_i)) + (1 - y_i) \log(1 - p(x_i)) \} \\
= - \sum_{i=1}^{n} \left\{ \frac{\sum_{j=1}^{n} \theta_j}{\sigma} (x_{ik} p_j (1 - p_i) - x_{ik} p_i (1 - p_j)) (p_j - p_i) K(p_j, p_i) \right\} y_i \log(p_i) + (1 - y_i) \log(1 - p_i).
The second term of Equation (11) is as follows:

\[- \sum_{i=1}^{n} \hat{w}(p(x_i)) \frac{\partial}{\partial \beta_k} \{y_i \log (p(x_i)) + (1 - y_i) \log (1 - p(x_i)) \} \]

\[= \sum_{i=1}^{n} \left( \sum_{j=1}^{n} \theta_j K(p_j, p_i) \right) (x_{ik} (y_i - p_i)). \]

The final term of Equation (11) is as follows:

\[- n \frac{\partial}{\partial \beta_k} \lambda \| \beta \|^2 = -2 n \lambda \beta_k. \]

3.2.2. Estimating \(w_a(p)\)

\(w_a(p)\) is modeled as in equation (10). Here, \(y_i\) and fitted value \(p_i\) are used to estimate \(\theta\) in equation (10). \(y_i\) is used as the observed value of \(p_i\). We can estimate \(\theta\) via relative unconstrained least-squares importance fitting (RuLSIF; Yamada et al. 2013), to minimize the following objective function:

\[L_\theta(\theta | \beta) = \frac{1}{2} \theta^T \hat{H} \theta - \hat{h}^T \theta + \frac{\lambda}{2} \| \theta \|^2, \tag{12} \]

where

\[(\hat{H})_{lm} = \frac{\alpha}{n} \sum_{i=1}^{n} K(\hat{p}_i, \hat{p}_l) K(\hat{p}_i, \hat{p}_m) + \frac{(1 - \alpha)}{n} \sum_{i=1}^{n} K(y_i, \hat{p}_l) K(y_i, \hat{p}_m), \]

\[(\hat{h})_{l} = \frac{1}{n} \sum_{i=1}^{n} K(\hat{p}_i, \hat{p}_l). \]

The tuning parameters \(\sigma\) and \(\lambda_\theta\) can be optimized via cross-validation.

Algorithm I: Estimation Algorithm

\begin{itemize}
  \item **Input:** Data \(\{x_i\}_{i=1}^{n}\) and True Label \(\{y_i\}_{i=1}^{n}\)
  \item **Output:** Parameter Vector \(\hat{\beta}\)
  \begin{enumerate}
    \item \(t = 0\)
    \item \(\beta^{(0)} = \text{minimize cross-entropy}(\{x_i\}_{i=1}^{n}, \{y_i\}_{i=1}^{n})\)
    \item \(\hat{p}_i = 1/(1 + \exp(-\beta^{(t-1)^T} x_i)) \quad (i = 1, 2, \ldots, n)\)
    \item Estimate \(\hat{\theta}\) via RuLSIF
    \item \(\hat{\beta}^{(t)} = \text{minimize weighted cross-entropy}(\hat{\theta}, \{x_i\}_{i=1}^{n}, \{y_i\}_{i=1}^{n})\)
    \item \(t = t + 1\)
    \item Repeat Steps 3-6 until \(\|\hat{\beta}^{(t-1)} - \hat{\beta}^{(t)}\|^2 < \epsilon\)
  \end{enumerate}
\end{itemize}

4. Experiment

We compare the proposed method, whose \(\alpha\) was set to 0.5, with L2 regularized logistic regression, L2 regularized logistic regression via SMOTE (Chawla et al. 2002), L2 regularized logistic regression via a cost-sensitive method (Bbahnsen et al. 2014) and Janché’s model (Jansche 2005) with respect to the area under the curve (AUC) of the testing data. AUC is defined as follows:

\[\text{AUC} = \frac{\sum_{j=1}^{n_p} \sum_{i=1}^{n_n} I(p_i^p > p_j^p)}{n_p n_n}, \tag{13} \]

\]
where \(I(\cdot)\) is an indicator function. \(p^p_i (i = 1, 2, \ldots, n_p)\) are predicted scores of positive instances and \(p^n_j (j = 1, 2, \ldots, n_n)\) are predicted scores of negative instances.

To tune the hyper-parameters of the proposed method and L2 regularized logistic regression, we used grid search. Specifically, we set the range of penalty terms \(\lambda_\theta\) and \(\lambda_\beta\) to \(\{0.01, 0.1, 1, 10\}\), the range of \(\sigma\) for the Gaussian kernel to \(\{0.01, 0.1, 1, 10\}\), and the range of the number of variables to \(\{1, 2, \ldots, d\}\). We selected the best AUC performance hyper-parameters via five-fold cross-validation.

### 4.1. Real data example

Table 3 gives an overview of the datasets. We used two types of datasets to evaluate performance: one with a data ratio of 2:1 between majority and minority class labels and one with a ratio of 9:1. The Yeast\(^1\) dataset (Nakai and Kanehisa 1992) and Breast Tissue (Jossinet 1996) were taken from the UCI Repository of Machine Learning Databases at [http://www.ics.uci.edu/~mlearn/MLRepository.html](http://www.ics.uci.edu/~mlearn/MLRepository.html) and the Oil dataset (Kubat, Holte, and Matwin 1998) was taken from Open ML (https://openml.org/).

We first eliminated objects with NA observations. Then, the datasets were randomly divided into training and testing data. The ratio of the number of objects in the training and testing datasets was 7:3. The data were then normalized to a zero mean and unit variance. We simulated 100 times.

Tables 4–6 show the mean and standard deviation (s.d.) of AUC score, F-measures, and accuracy, respectively. From Table 4, SMOTE has good performance in terms of AUC for the Oil and Yeast datasets. The proposed method shows the second highest score for the Oil and Yeast datasets. By contrast, the proposed method shows good performance for the Breast Tissue dataset. This is because the predicted score is estimated by the proposed method with greater accuracy. From Table 5, the proposed method shows good performance according to the F-measure for all data. This is because the proposed method can maximize the F-measure. Finally, from Table 6, logistic regression has the highest accuracy score for the Oil dataset. For the Yeast and Breast Tissue datasets, logistic regression shows the second highest score. The proposed method shows the highest accuracy for the Yeast and Breast Tissue datasets, because logistic regression can maximize accuracy. This shows that the proposed method has the highest F-measure, implying that a better discriminant result can be obtained by using the F-measure.

### 4.2. Numerical example

We generated synthetic data as follows, such that \(i = 1, 2, \ldots, 100\).

- \(X_i \sim N_d(0, I_d)\)
- \(p_i = \frac{1}{1 + \exp(-\beta \cdot X_i)}\)
- \(Y_i \sim B(1, p_i)\)

\(^1\)Because the Yeast dataset is not binary, we replaced the ME2 class with a positive class and other classes with negative ones.
Table 7 shows the setting of the true \( \beta \) and positive case ratio for each dataset. Test data were also created under these settings. We simulated 100 independent datasets.

Table 8 shows the AUCs for the numerical example. SMOTE has the best performance under settings 1, 5, and 6. By contrast, the proposed method exhibits the best performance under settings 2, 3, and 4 and second best performance under settings 1, 5, and 6. For class-imbalanced data, SMOTE shows good performance from the perspective of AUC. Table 9 presents the F-measures for the numerical example. The proposed method has the best performance of all settings. As such, for class-imbalanced data, the proposed method shows good performance from the perspective of using the F-measure. Table 10 gives the accuracy for the numerical example, with the proposed method and logistic regression exhibiting the best performances.
However, from Tables 8 and 9, the proposed method is not better at estimating the probability of $\hat{p}$ as it has a better performance for discrimination, but not better AUC scores.

### 5. Conclusions

In this paper, we propose an F-measure maximizing logistic regression by using the relative density ratio. We demonstrate that the proposed method has better performance than the other methods with respect to the F-measure, for class imbalanced real data. The results indicate that the proposed method is useful when a discriminant is present. By contrast, the AUC of the proposed method is not higher than that of SMOTE because the weights of objects near the decision boundary are higher than those in the other methods. That is, the order of estimated probability is unstable near the decision boundary.

Our future work will involve developing this theoretical approach and improving predictive performance. There are five problems to be tackled. First, the theoretical predictive performance of our model is unclear. In the experiment, our method is used on imbalanced data. However, it possesses unclear asymptotic properties. Second, the performance of the approximation with the
relative F-measure should be investigated. In the experiment, the prediction performance of our model is effective with respect to AUC. However, we also need to consider the performance of the approximation. Third, the properties of high dimensional data should be investigated (e.g., Nakayama 2020). In this paper, we assumed \( n > d \). However, in a high-dimension, low-sample-size setting, the properties of our method are not clear. Fourth, we should consider a full kernel-based method. By extending Kernel uLSIF (Kanamori, Suzuki, and Sugiyama 2012), we can estimate the relative density ratio. To apply this method, it would be necessary to consider the estimation of predictive models using logistic regression analysis in the reproducing kernel Hilbert space. This method is not interpretable. However, if we assume \( P(y|x) \) as a non-linear function of explanatory variables and focus on prediction, the full kernel-based method is useful. Therefore, applying full kernel-based methods is one of our future directions. Finally, we should demonstrate the appropriateness of the relationship of the density ratio of \( p \) and \( \hat{p} \), and of the F-measure. Overall, we obtained better discriminant performance by using the density ratio experimentally, but we did not obtain theoretically relevant results.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP22K17862.

Disclosure statement

No potential competing interest was reported by the authors.

ORCID

Masaaki Okabe http://orcid.org/0000-0001-6299-8150
Jun Tsuchida http://orcid.org/0000-0003-2349-2709
Hiroshi Yadohisa http://orcid.org/0000-0002-1408-2655

References

Bahnsen, A. C., D. Aouada, and B. Ottersten. 2014. Example-dependent cost-sensitive logistic regression for credit scoring. In 13th International Conference on Machine Learning and Applications, 263–9.
