RESEARCH ARTICLE
doi: 10.2306/scienceasia1513-1874.2023.s003

ScienceAsia 49S (2023): **68–77**

A modification of logistic regression with imbalanced data: F-measure-oriented Lasso-logistic regression

Bui T. T. My[a,b,*], **Bao Q. Ta**[c]

[a] Department of Mathematical Economics, Ho Chi Minh University of Banking, Ho Chi Minh City, 7000 Vietnam
[b] Faculty of Mathematics and Statistics, College of Technology and Design, UEH University, Ho Chi Minh City, 7000 Vietnam
[c] Department of Mathematics, International University, Vietnam National University, Ho Chi Minh City, 7000 Vietnam

[*]Corresponding author, e-mail: mybtt@hub.edu.vn

Received 15 Jan 2023, Accepted 31 Aug 2023
Available online 27 Dec 2023

ABSTRACT: Logistic regression (LR) is one of the most popular classifiers. However, LR cannot perform effectively on imbalanced data. There are two approaches to imbalanced data for LR, including resampling techniques and modifications to the log-likelihood function. These approaches improve performance measures of LR in some cases, but their effectiveness is not robust in general. In this paper, we propose a classifier called F-measure-oriented Lasso-Logistic Regression (F-LLR) to deal with imbalanced data. The base learner of F-LLR is Lasso-Logistic regression (LLR) which imposes the prior on the magnitude of parameters by a hyper-parameter λ. The optimal λ is determined by an adjustment of the cross-validation procedure which aims for the highest F-measure instead of the highest accuracy. F-LLR addresses imbalanced data by the combination of Under-sampling and Synthetic Minority Oversampling Technique (SMOTE) selectively based on the scores of the training data. The empirical study shows that F-LLR increases F-measure and KS as compared with LLR and the traditional balanced methods, such as the resampling techniques (Random Under-sampling, Random Over-sampling, and SMOTE) and the modifications to log-likelihood function (Ridge and Weighted likelihood estimation).

KEYWORDS: cross-validation, F-measure, ridge, smote

MSC2020: 62H30 68T10 91C20

INTRODUCTION

Recently, although machine learning and data-mining algorithms are penetrating into several real applications of classification, Logistic regression (LR), a traditional model, is still in favor by several authors [1–3]. There are two prominent reasons for that. Firstly, the output of LR is the samples' conditional probabilities of belonging to the interest class, which are reasonable references to classify the samples. Secondly, LR shows a transparent model for interpretation while most machine learning and data-mining models operate as a 'black box' process. However, LR has some problems. The interpretive power of LR is based on the statistically significant level of parameters which is closely relevant to p-value. Nevertheless, p-value has been recently criticized since its meaning is usually misunderstood [4]. Furthermore, in imbalanced circumstances where the minority class is the interest object, the parameter estimation of LR can be biased and the conditional probability of belonging to the interest class can be under-estimated [5, 6]. As a consequence, LR usually misclassifies the interest class on imbalanced data.

In the literature on LR with imbalanced data, there were two main groups of methods, which were linked to the algorithm-level approach. They were Weighted Likelihood Estimation (WLE) [7–9] and Penalized Likelihood Regression (PLR) [10, 11]. Most of them were designed to reduce the parameter estimation bias and the predicted probability bias, especially in small samples. However, WLE needs the prior information of two classes in the population which is usually unavailable. Besides, some methods of PLR, such as FIR [6], FLIC, and FLAC [10] are quite sensitive to initial values in the computation process of maximum likelihood estimation. Therefore, solving LR with imbalanced data should consider both data-level and algorithm-level approaches and not make the computation process complex.

To deploy the ability of interpretation of LR and solve the imbalanced issue, we propose a binary classifier named *F-measure-oriented Lasso-Logistic regression* (F-LLR). F-LLR utilizes Lasso Logistic regression (LLR) as a base learner and integrates algorithm-level and data-level approaches to deal with imbalanced data. Lasso is a penalized shrinkage estimator and a feature selection method without p-value. In Lasso, the hyper-parameter λ is set by a new procedure called F-CV which is an adjustment of the ordinary cross-validation procedure (CV). F-CV finds the optimal λ by maximizing the cross-validation F-measure instead of the cross-validation accuracy as the way of CV. The proposed classifier F-LLR has two computation stages. In the first stage, LLR based on F-CV is applied to get the scores of all samples. In the second stage, accord-

ing to the scores, under-sampling (US) and Synthetic Minority Oversampling Technique (SMOTE) [12] are respectively used to re-balance the data set. Next, LLR based F-CV is applied again on the balanced data set to get the final classification results. The proposed classifier F-LLR experimented on nine real imbalanced data sets and its performance measures (KS and F-measure) are higher than the traditional approaches to imbalanced data of LR.

The paper is organized as follows. The Related Works section reviews the general background involved with LR and imbalanced data. The Methodology section describes the proposed classifier, the empirical data sets, the performance measures, and the implementation protocol. The Results and Discussions section presents the testing performance measures and the results of a statistical test. The final section is the Conclusion.

RELATED WORKS

The paper only focuses on the binary classification. In an imbalanced data set, the label of the minority class is denoted by "1", which is also called positive class. The label of the other is "0" and called negative one.

Imbalanced data in classification

Data for classification is considered imbalanced if there is a class with much fewer quantities than the other. In most of real applications, the minority class is always the most crucial object due to heavy losses if misclassified. Meanwhile, common classifiers are usually designed to get the predicted results if the accuracy of the model is the greatest. That makes the biased prediction toward the majority class [13]. Especially, when an extreme imbalance occurs, the minority class is thought about noise, so they are ignored for the highest accuracy. In short, it has failed in realizing patterns of the crucial class and the accuracy is not a rational metric to evaluate the performance of classifiers on imbalanced data sets.

There are two popular approaches to imbalanced data. They are algorithm-level and data-level [14–16].

The algorithm-level involves modifications of algorithm classifiers, such as modifying the decision thresholds to optimize a specific evaluation measure, assigning weights to samples in training data, setting loss matrix, or integrating cost-sensitive error functions, and so on. This approach directly tackles the consequences of imbalanced data and does not change the distribution of training data. However, in some circumstances, the algorithm-level approach was disapproved. For example, it was unreasonable for the difference in the loss when miss-classifying positive samples and the negative, or the weights of samples in weighted methods [15].

In contrast, the data-level approach directly balances data sets by re-sampling techniques with over-sampling and under-sampling as two typical representatives. Most of them are non-heuristic, such as random over-sampling (ROS) and random under-sampling (RUS), so they are easy to apply and independent from the classifier algorithms [17]. The most popular techniques are RUS, ROS, and SMOTE which have own drawbacks. For instance, RUS may waste important information of the original data and ROS often causes over-fitting [18, 19]. SMOTE balances the quantities of the two classes by creating synthetic positive samples [12]. Though SMOTE can prevent an over-fitting model, SMOTE pushes the boundary between two classes into the space of the majority class or causes overlapping classes [20]. Furthermore, empirical studies agreed that the combination of over and under-sampling families could work better than ones in only a family [12, 21, 22].

Logistic regression with imbalanced data

Logistic regression

Let $Y \in \{0,1\}$ be variable for labels and $X = (X_1,\ldots,X_p) \in \mathbb{R}^p$ be predictor variables. Logistic regression (LR) model is described as follows.

$$\pi(x) = P(Y=1|X=x) = \frac{e^{\beta_0+\beta x^T}}{1+e^{\beta_0+\beta x^T}}, \qquad (1)$$

where $\beta = (\beta_1,\ldots,\beta_p)$ and β_0 are the parameters showing the effects of the predictors on the conditional probability $\pi(x)$ which is also called the *score* of $x \in \mathbb{R}^p$.

Consider a sample data set of n independent observations: $\{(x_i,\ y_i) \in \mathbb{R}^{p+1},\ i = \overline{1,n}\}$, where $x_i \in \mathbb{R}^p$ is the vector expressing p features and $y_i \in \{0,1\}$ is the label of observation i-th. Then, the parameters in (1) can be estimated by maximizing the log-likelihood function:

$$\log \mathbb{L}(P(Y|X,\beta)) = \sum_{i=1}^{n} [y_i \log(\pi(x_i)) + (1-y_i)\log(1-\pi(x_i))]. \qquad (2)$$

The solution for (2) is computed by an interactive algorithm, such as Newton-Raphson method. A new observation x^* will be classified into the positive class if and only if its score is greater than a given threshold. We refer to [23] for a detailed discussion.

In imbalanced data, the parameter estimation of LR from (2) can be biased and the scores can be under-estimated [6]. Therefore, LR model usually misclassifies the positive samples.

Logistic regression with imbalanced data

There were two main groups of methods that focused on the intrinsic computation process of LR to reduce

the bias. They were Weighted Likelihood Estimation (WLE) and Penalized Likelihood Regression (PLR).

Weighted Likelihood Estimation (WLE) considers the weighted log-likelihood function:

$$\log \mathbb{L}_W\left(P\left(Y|X,\beta\right)\right) = \sum_{i=1}^{n} w_i \left[y_i \log\left(\pi\left(x_i\right)\right) + \left(1-y_i\right)\log\left(1-\pi\left(x_i\right)\right)\right]. \quad (3)$$

In (3), $w_i = \frac{\tau}{\overline{y}} y_i + \frac{1-\tau}{1-\overline{y}}\left(1-y_i\right)$ is the weight of the observation i-th in the sample data. Where τ and $\overline{y}$ are the proportions of the positive class in the population and in the sample, respectively. Details of WLE and extensive studies can be found in [5, 9, 24]. The obstacle of WLE is that the population proportion τ is unavailable.

Penalized Likelihood Regression (PLR) has the general form as follows:

$$\log \mathbb{L}^*\left(P\left(Y|X,\beta\right)\right) = \log \mathbb{L}\left(P\left(Y|X,\beta\right)\right) + A(\beta). \quad (4)$$

In (4), the term of $A(\beta)$ could be:

- Ridge-type: $A(\beta) = -\lambda \sum_{j=1}^{p} \beta_j^2$, where $\lambda > 0$ (see [9, 11]).
- Lasso-type: $A(\beta) = -\lambda \sum_{j=1}^{p} \left|\beta_j\right|$, where $\lambda > 0$ (see [25, 26]).
- Firth-type: $A(\beta) = \frac{1}{2}\log(\det(I(\beta)))$, where $I(\beta)$ is the Fisher information matrix (see [6]).

Regarding Firth-type (FIR), although the parameter estimation bias can be reduced, it introduces the bias in the scores. To overcome this drawback, FLIC and FLAC, which are the modification of FIR, are proposed. Although FLIC and FLAC perform better than FIR, they cannot win Ridge on most empirical and simulation data sets (see [10]). Besides, FIR, FLIC, and FLAC were quite sensitive to initial values in the computational process of the maximum likelihood estimation.

Regarding Ridge and Lasso, the penalty parameter λ controls the magnitude of the estimations of $\beta_j\,(j \neq 0)$ (denoted $\widehat{\beta_j}$) which are found by Coordinate descent algorithm [27]. The optimal λ can be usually determined by cross-validation procedure (CV) based on the default threshold of 0.5 and minimizing the cross-validation Error rate (or maximizing the cross-validation Accuracy). Ridge may lead to a dense estimation of β, which has very few values zero of $\widehat{\beta}$. In high dimension data, Ridge takes a large interval of computation time. Analogous to Ridge, Lasso is a penalized shrinkage estimator. Besides, Lasso is also a feature selection method without any p-value. In Lasso, the larger λ is, the more number of $\widehat{\beta_j}$ are zero. Thus, Lasso retains only the predictors closely relevant to the response. The details of Lasso and Ridge can be found in [28]. However, Lasso does not deal with imbalanced data. Some studies applied SMOTE to rebalance data before performing Lasso (see [29, 30]). Despite the fact that, SMOTE causes the overlapping classes which decreases the performance measures of classifiers [20].

METHODOLOGY

Reviewing the literature on LR with imbalanced data leads to some conclusions. LR can still employ the ability of interpretation with the penalized version of Lasso. With imbalanced data, it should be considered the hybrid of both intrinsic (algorithm-level) and extrinsic (data-level) algorithms of Lasso Logistic regression (LLR). Moreover, the data-level approach should be examined to boost the advantages and restrict the disadvantages of the re-sampling techniques. For example, SMOTE should be only applied to the safe subset of the minority class which consists of samples with typical characteristics of the interest class. In addition, the algorithm-level approach can be used to modify the computation process of LR and can support the application of re-sampling techniques.

Inspired by the idea of the hybrid approach to LR with imbalanced data, the paper proposed a modification of LR named F-measure-oriented Lasso-Logistic regression (F-LLR).

The proposed classifier

F-LLR utilizes Lasso Logistic regression (LLR) as a base learner. Instead of using CV to find the optimal λ, a modification of CV, called F-measure-oriented cross-validation (F-CV), is proposed. In F-CV, the criterion to evaluate the optimal λ is F-measure, a more suitable metric than Accuracy on imbalanced data. The details of CV-F are described in Table 1 and illustrated in Fig. 1.

Under the notations Table 1, with every threshold α_j, the cross-validation F-measure, F_{ij}, is an estimate of the testing F-measure of the classifier LLR(λ_i). When the penalty parameter λ and the threshold α take all values in the series $\{\lambda_i\}_1^h$ and $\{\alpha_j\}_1^l$, respectively, F_{i0j0} determined at Step 10 is an estimate of the highest testing F-measure of LLR(λ) on data set T. Therefore, F-CV indicates not only the optimal penalty parameter λ_{i0} but also the optimal threshold α_{j0} which correspond to F_{i0j0}.

The proposed classifier F-LLR has two computation stages. In the beginning, all of the samples of the training data are scored by F-CV. Then, according to the samples' scores, Under sampling (US) and SMOTE are respectively applied to balance the training data set. Finally, on the balanced data set, LLR based F-CV builds a classifier F-LLR. The computation process for F-LLR is shown in Table 2.

Table 1 F-measure-oriented cross-validation procedure.

Input:	A training data set T, a series of $\{\lambda_i\}_1^h$, a series of threshold $\{\alpha_j\}_1^l$, and an integer K.
1.	Randomly divide T into K equal- sized subsets: $T_1, \ldots, T_K$.
2.	For $i \in \{1, 2, \ldots, h\}$, do following:
3.	For $j \in \{1, 2, \ldots, l\}$ do following:
4.	For $k \in \{1, 2, \ldots, K\}$ do following:
5.	On the $T \backslash T_k$, apply LLR with λ_i to get a fitted model LLR(λ_i).
6.	On the T_k, apply LLR(λ_i) to get the scores of the samples of T_k.
7.	Compare the scores with the threshold α_j to get the labels of T_k.
8.	Calculate F-measure, denoted F_{ijk}.
9.	$F_{ij} = \frac{1}{K}\sum_{k=1}^{K} F_{ijk}$.
10.	$F_{i_0 j_0} = \max_{i,j}\{F_{ij}\}$
Output:	The classifier LLR(λ_{i_0}), the optimal penalty λ_{i_0}, and the optimal threshold α_{j_0}.

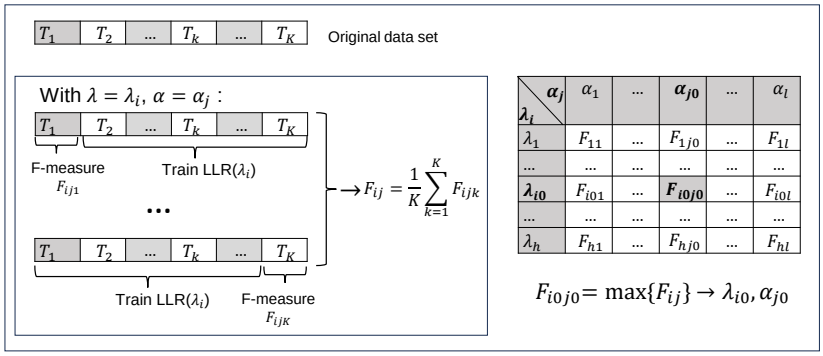

Fig. 1 Illustration for F-CV.

The combination of US and SMOTE aims to remove the useless samples and increase the useful ones. The fact that the higher the score of negative samples is, the greater the chance of being misclassified is. Those may be noise, borderline, or overlapping samples which decrease the performance measures of the classifiers [19]. Thus, US is utilized to eliminate a proportion of the negative class which contains the upper high-scored samples. Next, instead of applying SMOTE to the whole minority class, SMOTE just performs on the subset consisting of the positive samples with high scores. The idea contrasts with the application of US. The high-scored positive samples are usually identified correctly across thresholds. This practice is meant to emphasize the prominent characteristic of the positive which is useful for the identification of the

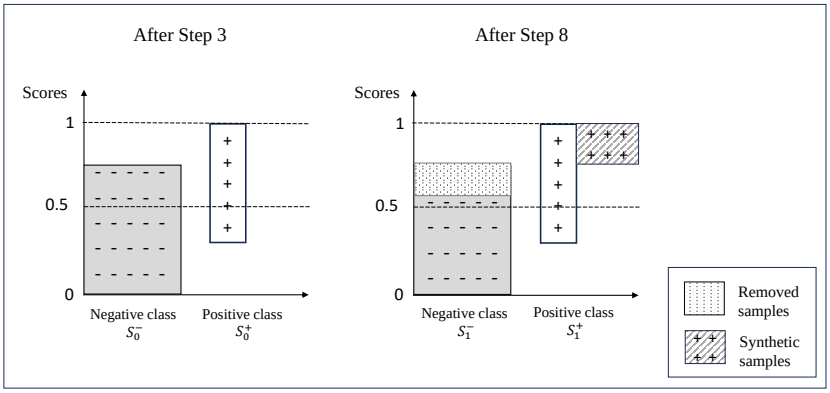

Fig. 2 Illustration for F-LLR.

minority class. Furthermore, the high-scored positive samples are in the safe region that is usually far from the borderline, so applying SMOTE here can prevent overlapping classes. Fig. 2 shows the ideas under the steps of F-LLR classifier.

Data for empirical study

Credit scoring is a typical example of imbalanced classification since the number of bad customers is always far less than the number of good ones. Eight credit-scoring data sets are used in the experimental study. They are Australian data (AUS), German data (GER), Taiwanese data (TAI), Credit risk data (Credit 1), Credit default data (Credit 2), Credit Card data (Credit 3), Bank personal loan data (Credit 4), and Vietnamese data (VN). Moreover, a data set of hepatitis patients (Hepa), which is not only imbalanced but also has a small size of the positive class, is also investigated.

All data sets suffer imbalanced status with different levels evaluated by the imbalanced ratio (IR), which is the rate of the quantity of the negative and positive classes. The details of the data sets are presented in Table 3 in the order of IR (shown in the column named 'IR') from the smallest to the highest. The first group of data sets, including AUS, GER, TAI, and Credit 1, are imbalanced data at a low level ($IR \leqslant 5$). AUS, GER, and TAI data sets publicized on the UCI machine learning repository are familiar with credit scoring studies. Besides, Credit 1 is the subset randomly drawn from the original data set at the rate of 20% to save computation time. Credit 1 still maintains the same IR as the original data on the Kaggle website. The second group consisting of Credit 2, 3, 4, and Hepa suffers average imbalanced status ($5 < IR \leqslant 10$). Credit 3 is formed in a similar way to Credit 1 but at the rate of 10%. The last group is Vietnamese data collected from a commercial bank in Vietnam in the period 2019 - 2020. This is the most severely imbalanced data among experimental data sets. Except for the last data set, eight others are public data on the UCI library and Kaggle website with transparent sources indicated in Table 3.

All observations with missing values of features are omitted. Moreover, all numeric features of data sets are standardized to have zero mean and unit deviation.

Performance measures

Accuracy is not a reasonable measure for classifiers on imbalanced data [13]. Instead, AUC, KS, F-measure, and G-mean are utilized to evaluate the performance of classifiers considered in the paper. Among them, AUC and KS are free-threshold measures that judge the general effectiveness of classifiers. Meanwhile, F-measure and G-mean depend on the threshold which is a reference value to distinguish 'positive' and 'negative'. Details of AUC, KS, F-measure, and G-mean can be found in the documents [31, 32].

About F-measure, it is the harmonic mean of Precision and Recall as the formula (5). Precision is the ratio of true positive samples among the predicted positive and Recall is the proportion of the predicted positive samples in the positive class. F-measure is high if and only if both Precision and Recall are high. On imbalanced data sets, LR and LLR usually give a high Precision and low Recall. It means that few of the positive samples are classified correctly. On the contrary, when boosting the Recall but ignoring the Precision of imbalanced data, it leads to an extreme classifier that cannot identify the negative. The bias toward Precision or Recall can cause unnecessary losses, for example in credit scoring or medical diagnosis field. Therefore, in the procedure to find the optimal λ of LLR, the highest F-measure is a more reasonable target than the highest Accuracy.

$$\text{F-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{5}$$

In this paper, the F-measure and G-mean corresponding to threshold α is denoted by F-measure(α) and G-mean(α).

Implementation protocol

The performance of F-LLR is compared with popular balanced methods of LLR, such as LLR with a resampling technique (RUS, ROS, or SMOTE) and Ridge. Especially, on Vietnamese data set, F-LLR was compared with WLE due to the available value of τ ($\tau = 1.7\%$) which is the bad debt ratio in the Vietnamese banking system in the period 2019-2020. Ridge-LR and WLE, the representative of LR with the algorithm-level approach, are chosen to compare with F-LLR since these methods worked better than others according to previous studies [5, 10]. The optimal λs of the models, including LLR, RUS-LLR, ROS-LLR, SMOTE-LLR, and Ridge, are determined by the original version of CV.

The general implementation protocol is described in Table 4. Furthermore, we set up the series of hyper-parameters as follows:

- The series of lambdas $\{\lambda_i\}_{i=1}^{100}$ consists of 100 equal-distanced values from 0.005 to 0.0001.
- The series of thresholds $\{\alpha_j\}_{j=1}^{50}$ consists of 50 equal-distanced values from 0.01 to 0.7. We choose 0.7 as the upper bound of the series of thresholds because if the threshold for distinguishing two classes is too high, there are many positive samples misclassified.
- The series of rates for under-sampling $\{r_U\}_1^{20}$ consists of 20 values from 0.05 to $0.5\times(\text{IR}-1)/\text{IR}$ and satisfying $(1-r_U)|S_0^-| > |S_0^+|$. If RUS is applied, a number of negative samples which account for $(\text{IR}-1)/\text{IR}$ of the negative class will

Table 2 The procedure for F-LLR classifier.

Input:	Training data set $T_0 = S_0^+ \cup S_0^-$, where S_0^+ and S_0^- are the positive and negative class. Series of penalties $\{\lambda_i\}_1^h$, series of thresholds $\{\alpha_j\}_1^l$, and an integer K. r_U: a rate for US; and r_S: a rate for SMOTE (satisfying $(1-r_U)\|S_0^-\| > \|S_0^+\|$).
Stage 1	
1.	Apply F-CV on T_0 to get the classifier LLR(λ_0).
2.	Apply LLR(λ_0) to score all samples of T_0.
Stage 2	
3.	Order the samples of S_0^+ and S_0^- by their scores from the highest to the lowest.
4.	On S_0^-, remove $(r_U \times \|S_0^-\|)$ upper high-scored samples to get S_1^-.
5.	Determine the subset of S_0^+ consisting of $(r_S \times \|S_0^+\|)$ upper high-scored samples called S_0^{++}.
6.	$m = \dfrac{\|S_1^-\| - \|S_0^+\| \times (1-r_S)}{\|S_0^+\| \times r_S}$
7.	Apply SMOTE on S_0^{++} to create $(m-1)r_S \times \|S_0^+\|$ synthetic samples.
8.	The new positive class S_1^+.
9.	Apply F-CV on the balanced training set $T_1 = S_1^+ \cup S_1^-$.
Output:	The classifier LLR(λ_1) and the optimal threshold α_1.

$|A|$ denotes the quantity of the data set A.

Table 3 The description of the experimental data sets.

Data sets	Size	# positive[a]	IR	# features[b]	# num features[c]	Source
AUS	690	307	1.25	14	6	
GER	1,000	300	2.33	19	7	ics.uci.edu/ml/datasets
TAI	30,000	6,636	3.52	23	14	
Credit 1	5,752	1,237	3.65	11	7	(1)
Credit 2	9,709	1,283	6.57	18	5	(2)
Credit 3	12,600	1,525	7.26	11	5	(3)
Hepa	589	63	8.35	12	11	ics.uci.edu/ml/datasets/hepatitis
Credit 4	5,000	480	9.42	11	6	(4)
VN	10,889	602	17.09	12	0	Vietnam

[a]: the number of positive class; [b]: the number of features; [c]: the number of numeric features.
(1): https://www.kaggle.com/datasets/laotse/credit-risk-dataset.
(2): https://www.kaggle.com/datasets/gargvg/univai-dataset.
(3): https://www.kaggle.com/datasets/samuelcortinhas/credit-card-classification-clean-data.
(4): https://www.kaggle.com/datasets/teertha/personal-loan-modeling.

be removed randomly. However, in this employment, we do not eliminate too many negative samples to restrict the loss of valuable information from the majority class.

- The series of rates for SMOTE $\{r_S\}_1^{20}$ consists of 20 values from 0.05 to 0.75. Usually, SMOTE balances data by using 100% samples of the minority class to generate synthetic samples in their neighborhoods. In our way, the upper bound rate for SMOTE is 0.75 since we focus on the top-scored positive samples to restrict the overlapping issues, a typical drawback of the standard SMOTE.

Note that on each data set, the considered classifiers carry out 20 times and the performance measures of 20 times are averaged to get the robust results.

RESULTS AND DISCUSSION

Results

With F-LLR, a minor experiment on some values of r_U (the rate for US) and r_S (the rate for SMOTE) suggested that the optimal value of r_U was in the range $[0.05; 0.25]$ while the one of r_S was in $[0.20; 0.75]$. Where, the Hepa data set, which has a very small positive class and suffers average imbalanced level, performs at the $r_U = 0.07$ and $r_S = 0.75$. It can be implied that SMOTE technique is prioritized to US in the protocol of F-LLR on experimental data.

The average performance measures of considered classifiers were recorded in Table 5. In comparison to other classifiers, F-LLR showed better performance on experimental data sets. On TAI, Credit 2, Credit 3, and Hepa, F-LLR was the most prominent classifier since F-LLR won the other at least three performance metrics. On other data sets (except Credit 4), F-LLR

Table 4 The implemetation protocol.

Steps	Contents
1.	Set up the series: $\{\lambda_i\}_1^{100}$, $\{\alpha_j\}_1^{50}$, $\{r_{Uu}\}_{u=1}^{20}$, and $\{r_{Ss}\}_{s=1}^{20}$.
2.	Split randomly the data set into the training and testing set (70% : 30%).
On the training set:	
3.	For $u \in \{1,\ldots,20\}$ do following:
4.	For $s \in \{1,\ldots,20\}$ do following:
5.	Apply the Algorithm F-LLR$\left(\{\lambda_i\}_1^{100}, \{\alpha_j\}_1^{50}, r_{Uu}, r_{Ss}\right)$ to build the classifiers F-LLR.
6.	Determine the optimal threshold α_j^* of F-LLR.
7.	Determine the optimal F-LLR based on the highest F-measure across r_{Us} and r_{Ss}, where $u,s \in \overline{1,20}$.
8.	Build the classifiers LLR, RUS-LLR, ROS-LLR, SMOTE-LLR, and Ridge, respectively.
9.	For each classifier, determine the optimal threshold α_j^* corresponding to the highest F-measure(α_j^*).
	On the training set of Vietnamese data, run above ones and WLE.
On the testing set:	
10.	Apply the considered classifiers, respectively.
11.	Calculate AUC, KS, F-measure(α_j^*), and G-mean(α_j^*) of all considered classifiers.
12.	Repeat from Step 2 to Step 11 twenty times.
13.	Average twenty values of AUC, KS, F-measure, and G-mean.

outperformed by two metrics.

Especially, on VN – the most imbalanced data, F-LLR was the most prominent classifier. F-LLR displayed the highest KS and F-measure. Besides, ROS-LLR won against F-LLR in G-mean and WLE won in AUC. Despite the fact that, the difference in G-mean between F-LLR and ROS-LLR was not significant. It was similar to the difference in AUC between F-LLR and WLE. Furthermore, Ridge-LR worked worse than all considered classifiers.

In general, F-LLR showed the highest KS and F-measure across data sets. In contrast, data-level approach could not deal with imbalanced data on GER, Credit 2, and Hepa. The techniques ROS, RUS, and SMOTE even decreased the performance measures of the LLR classifier. Besides, Ridge-LR seemed a competitor with F-LLR in some cases, such as GER and Credit 4.

About the optimal thresholds of LLR, they are quite higher than the ones of the other classifiers.

Statistical test

To have a confident conclusion on the effectiveness of F-LLR, Sign test was utilized. This test does not assign any assumption of the distribution of performance measures. It just counts the number of data sets on which the interest classifier wins the others. Details of Sign test can be found in [33]. When comparing multiple classifiers, it can be performed pairwise comparisons and record the results in a matrix. When considering the interest and another classifier, there are two possibilities: win or not. Thus, the number of wins follows the binomial distribution Binorm(N,p). Under the null hypothesis that is the two classifiers are equivalent, the parameters of this distribution are:

- N: the number of the empirical data sets.
- $p = 0.5$: the probability of winning under the null hypothesis.

With the Binorm$(N, 0.5)$, the critical number of wins can be calculated. For example, with $N = 9$, the critical values at the significant level of $\alpha = 5\%$ (or 10%) is $w_\alpha = 8$ (or $w_\alpha = 7$) [34]. It means the interest classifier is significantly better than another if it performs better on at least w_α data sets.

According to the results in Table 5, we organized two tests which were overall and pairwise comparisons to conclude the effectiveness of F-LLR. However, it could not be compared the performance measures of F-LLR and WLE since there was only one observation of this comparison. The number of wins of F-LLR were shown in Table 6. F-LLR won the others seven times by KS and F-measure in overall comparison. That implied the KS and F-measure of F-LLR were significantly higher than the ones of others at the level of 10% in overall comparison. In pairwise ones, there were some notes:

- By AUC: F-LLR won LLR and RUS-LLR on all nine data sets while it won ROS-LLR on five, SMOTE-LLR on six, and Ridge-LLR on six data sets. Therefore, F-LLR was only significantly better than LLR and RUS-LLR.
- By G-mean: F-LLR won LLR on eight data sets. Besides, F-LLR won RUS-LLR, SMOTE-LLR, and Ridge-LLR on seven data sets but won ROS-LLR on six ones. Thus, except for ROS-LLR, F-LLR significantly won the others.

Discussion

According to the empirical study and statistical test, the proposed classifier F-LLR with the combination of US and SMOTE under the control of the samples' scores completely beat RUS-LLR (by 4 performance

Table 5 The average testing performance measures of classifiers.

Data sets	Measures	LLR	RUS-LLR	ROS-LLR	SMOTE-LLR	Ridge-LR	WLE	F-LLR
AUS	AUC	.9221	.9238	.9246	.9236	.9276	——	.9245
	KS	.7450	.7503	.7546	.7559	.7530	——	.7525
	F-measure	.8464	.8483	.8516	.8527	.8534	——	.8541
	G-mean	.8578	.8589	.8624	.8642	.8565	——	.8645
	Threshold*	.4825	.4552	.4954	.4968	.3086	——	.5066
GER	AUC	.7722	.7646	.7634	.7623	.7845	——	.7782
	KS	.4605	.4489	.4552	.4551	.4618	——	.4627
	F-measure	.5948	.5643	.5723	.5708	.5968	——	.5987
	G-mean	.7070	.6534	.6726	.6694	.7127	——	.7097
	Threshold*	.3428	.3793	.3834	.3786	.2725	——	.4337
TAI	AUC	.6137	.7210	.7224	.7232	.6139	——	.7221
	KS	.3229	.3798	.3804	.3802	.3230	——	.3859
	F-measure	.4356	.3958	.4008	.4060	.4355	——	.4966
	G-mean	.5628	.4624	.4756	.4884	.5622	——	.6763
	Threshold*	.2750	.3173	.3266	.3322	.2750	——	.4269
Credit 1	AUC	.8523	.8528	.8532	.8529	.8524	——	.8529
	KS	.5646	.5660	.5649	.5637	.5655	——	.5662
	F-measure	.6262	.6166	.6156	6135	.6242	——	.6293
	G-mean	.7582	.7767	.7754	.7742	.7754	——	.7653
	Threshold*	.3259	.5500	.5500	.5500	.2741	——	.5691
Credit 2	AUC	.5758	.5741	.5743	.5752	.5309	——	.5761
	KS	.1744	.1678	.1701	.1706	.1281	——	.1748
	F-measure	.2832	.2677	.2703	.2712	.2632	——	.2802
	G-mean	.5428	.2918	.2226	.3036	.4250	——	.5838
	Threshold*	.1190	.3940	.4233	.4224	.1310	——	.3414
Credit 3	AUC	.5853	.5827	.5872	.5874	.5777	——	.5876
	KS	.1542	.1519	.1600	.1596	.1463	——	.1602
	F-measure	.2597	.2559	.2563	.2557	.2574	——	.2603
	G-mean	.5039	.2291	.2045	.2457	.4989	——	.5539
	Threshold*	.1110	.2712	.3822	.3539	.1148	——	.3667
Hepa	AUC	.9252	.8975	.9228	.9238	.9442	——	.9480
	KS	.7967	.7570	.8325	.8166	.8620	——	.8563
	F-measure	.7564	.6215	.7529	.7242	.7529	——	.7717
	G-mean	.7913	.8224	.9052	.8554	.7900	——	.9059
	Threshold*	.2622	.3442	.3968	.3757	.1959	——	.2658
Credit 4	AUC	.9409	.9435	.9436	.9405	.9560	——	.9462
	KS	.8355	.8671	.8625	.8682	.8430	——	.8691
	F-measure	.6418	.5224	.5295	.5254	.6428	——	.5335
	G-mean	.7677	.8180	.8253	.8235	.7715	——	.7923
	Threshold*	.3086	.3603	.3577	.3500	.2914	——	.3416
VN	AUC	.7885	.7946	.8019	.7996	.7883	.8065	.7956
	KS	.5429	.5462	.5496	.5505	.5042	.5510	.5584
	F-measure	.3111	.2594	.2661	.2793	.2754	.2756	.3406
	G-mean	.7252	.7633	.7657	.7106	.6290	.6306	.7653
	Threshold*	.1400	.4200	.4600	.3800	.1250	.0875	.4175

*: The optimal threshold corresponding to the highest trained F-measure. The underlined values is the highest in each row.

measures) and SMOTE-LLR (by 3 ones). Furthermore, on nine real data sets, F-LLR outperformed other considered classifiers in KS and F-measure. That meant F-LLR could separate the true positive distribution and the false positive distribution better than the others. Moreover, F-LLR showed the best trade-off between Precision and Recall. In credit scoring application, Recall is more important than Precision since the bad customers are always the crucial objects to identify in a credit scoring process. However, if classifiers stress on Recall and ignore Precision, a large number of good customers are rejected. This is also an unpleasant

Table 6 The number of wins of F-LLR on experimental data sets.

Performance measures	Pairwise comparisons					Overall comparison
	LLR	RUS-LLR	ROS-LLR	SMOTE-LLR	Ridge-LLR	
AUC	9	9	5	6	6	3
KS	9	8	8	8	7	7
F-measure	7	9	9	9	8	7
G-mean	8	7	6	7	7	4

With $N = 9$, the critical values $w_{0.05} = 8, w_{0.1} = 7$.

scenario for the financial organizations which operate for profit. Therefore, F-LLR with the remarkable ability to boost F-measure is an effective classifier for credit scoring.

CONCLUSION

LR is a very popular traditional classifier though there are many modern models available today. Similar to common classifiers, LR works ineffectively on imbalanced data sets. The algorithm-level and data-level approaches cannot increase the performance measures of LR in many imbalanced data sets. Consequently, the application fields of LR can be narrowed in spite of its strengths.

Taking advantage of the LLR, algorithm-level, and data-level approach, the proposed classifier F-LLR prepared a balanced training set by removing unnecessary negative samples and increasing essential positive samples by targeted applying US and SMOTE. Besides, the optimal penalty parameter λ and the optimal threshold of LLR were determined by a new procedure called F-CV, an adjustment of the ordinary CV. The modifications to the intrinsic algorithm of LLR and the re-sampling techniques made F-LLR more effective in KS and F-measure than the original versions, such as LLR, RUS-LLR, ROS-LLR, SMOTE-LLR, Ridge-LR, and WLE. This opens up the possibility of applying LLR on fields with severely imbalanced data while it is necessary to identify the input features which affects significantly on the classification results, for example credit fraud detection or cancer diagnosis. However, the best values of the hyper-parameters r_U and r_S of F-LLR should be investigated deeply by experiments. Besides, F-LLR should be applied to more real data sets of other fields to have a robust conclusion of its effectiveness.

Acknowledgements: The research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number C2021-34-02.

REFERENCES

1. Bektas J, Ibrikci T, Ozcan IT (2017) Classification of real imbalanced cardiovascular data using feature selection and sampling methods: a case study with neural networks and logistic regression. *Int J Artif Intell Tools* **26**, 1750019.
2. Khemais Z, Nesrine D, Mohamed M (2016) Credit scoring and default risk prediction: A comparative study between discriminant analysis & logistic regression. *Int J Econ Finance* **8**, 39–53.
3. Muchlinski D, Siroky D, He J, Kocher M (2016) Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Political Anal* **24**, 87–103.
4. Goodman S (2008) A dirty dozen: twelve *p*-value misconceptions. *Semin Hematol* **45**, 135–140.
5. King G, Zeng L (2001) Logistic regression in rare events data. *Political Anal* **9**, 137–163.
6. Firth D (1993) Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38.
7. Manski CF, Lerman SR (1977) The estimation of choice probabilities from choice based samples. *Econometrica* **45**, 1977–1988.
8. Ramalho EA, Ramalho JJ (2007) On the weighted maximum likelihood estimator for endogenous stratified samples when the population strata probabilities are unknown. *Appl Econ Lett* **14**, 171–174.
9. Maalouf M, Trafalis TB (2011) Robust weighted kernel logistic regression in imbalanced and rare events data. *Comput Stat Data Anal* **55**, 168–183.
10. Puhr R, Heinze G, Nold M, Lusa L, Geroldinger A (2017) Firth's logistic regression with rare events: accurate effect estimates and predictions? *Stat Med* **36**, 2302–2317.
11. Park MY, Hastie T (2008) Penalized logistic regression for detecting gene interactions. *Biostatistics* **9**, 30–50.
12. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: synthetic minority over-sampling technique. *J Artif Intell Res* **16**, 321–357.
13. Brown I, Mues C (2012) An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Syst App* **39**, 3446–3453.
14. Ertekin S, Huang J, Bottou L, Giles L (2007) Learning on the border: active learning in imbalanced data classification. In: *Proc of the 16th ACM Conf on Conference on Information and Knowledge Management*, Lisbon, pp 127–136.
15. Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H, Bing G (2017) Learning from class-imbalanced data: Review of methods and applications. *Expert Syst Appl* **73**, 220–239.
16. Xiao J, Zhou X, Zhong Y, Xie L, Gu X, Liu D (2020) Cost-sensitive semi-supervised selective ensemble model for customer credit scoring. *Knowl Based Syst* **189**, 105118.
17. López V, Fernández A, García S, Palade V, Herrera F (2013) An insight into classification with imbalanced data: Empirical results and current trends on using data

intrinsic characteristics. *Inf Sci* **250**, 113–141.
18. Barandela R, Valdovinos RM, Sánchez JS (2003) New applications of ensembles of classifiers. *Pattern Anal Appl* **6**, 245–256.
19. Batista GE, Prati RC, Monard MC (2004) A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explor News* **6**, 20–29.
20. Bellinger C, Drummond C, Japkowicz N (2016) Beyond the boundaries of smote. In: *Machine Learning and Knowledge Discovery in Databases*, Riva del Garda, pp 248–263.
21. Shamsudin H, Yusof UK, Jayalakshmi A, Khalid MNA (2020) Combining oversampling and undersampling techniques for imbalanced classification: A comparative study using credit card fraudulent transaction dataset. In: *2020 IEEE 16th International Conference on Control & Automation (ICCA)*, Singapore, pp 803–808.
22. Revathi M, Ramyachitra D (2021) A modified borderline smote with noise reduction in imbalanced datasets. *Wirel Pers Commun* **121**, 1659–1680.
23. James G, Witten D, Hastie T, Tibshirani R (2013) *An Introduction to Statistical Learning*, 2nd edn, Springer New York, New York.
24. Xie Y, Manski CF (1989) The logit model and response-based samples. *Sociol Methods Res* **17**, 283–302.
25. Li Y, Yu C, Qin Y, Wang L, Chen J, Yi D, Shia BC, Ma S (2015) Regularized receiver operating characteristic-based logistic regression for grouped variable selection with composite criterion. *J Stat Comput Simul* **85**, 2582–2595.
26. Fu GH, Xu F, Zhang BY, Yi LZ (2017) Stable variable selection of class-imbalanced data with precision-recall criterion. *Chemom Intell Lab Syst* **171**, 241–250.
27. Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* **33**, 1–22.
28. Hastie T, Tibshirani R, Wainwright M (2015) *Statistical Learning with Sparsity: The Lasso and Generalizations*, 1st edn, Chapman & Hall/CRC, New York.
29. Kitali AE, Alluri P, Sando T, Wu W (2019) Identification of secondary crash risk factors using penalized logistic regression model. *Transp Res Rec* **2673**, 901–914.
30. Shrivastava S, Jeyanthi PM, Singh S (2020) Failure prediction of Indian banks using smote, lasso regression, bagging and boosting. *Cogent Econ Finance* **8**, 1729569.
31. Josephine SA (2017) Predictive accuracy: A misleading performance measure for highly imbalanced data classified negative. In: *SAS Global Forum*, Orlando, pp 1–12.
32. Řezáč M, Řezáč F (2011) How to measure the quality of credit scoring models. *Finance a úvěr: Czech J Econ Finance* **61**, 486–507.
33. Sheskin DJ (2003) *Handbook of Parametric and Nonparametric Statistical Procedures*, 3nd edn, Chapman and Hall/CRC, New York.
34. Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res* **7**, 1–30.