Robust Bounded Logistic Regression in the Class Imbalance Problem

Guibiao Xu*, Bao-Gang Hu* and Jose C. Principe†
*NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China
Email: xuguibiao@gmail.com, hubg@nlpr.ia.ac.cn
†CNEL, Department of Electrical & Computer Engineering, University of Florida, Gainesville, FL, 32611, USA
Email: principe@cnel.ufl.edu

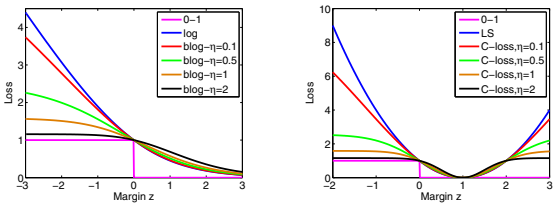
***Abstract*—In this paper, we propose to deal with the problems of logistic regression with outliers and class imbalance, which are common in a wide range of practical applications. The robust bounded logistic regression with different error costs is developed to reduce the combined influence of outliers and class imbalance. First, inspired by the Correntropy induced loss function, we develop the bounded logistic loss function which is a monotonic, bounded and nonconvex loss and thus robust to outliers. With the bounded logistic loss, we construct a new robust logistic regression. Second, under the principle of cost-sensitive learning, we assign different error costs for different classes in order to reduce the sensitiveness of the new robust logistic regression to class imbalance. Using the half-quadratic optimization method, it is easy to optimize the proposed logistic regression model. Experimental results demonstrate that our proposed method improves the performance of logistic regression on the datasets with outliers and class imbalance.**

I. Introduction

Logistic regression is a popular classifier in data mining and machine learning, and has been shown to be competitive with other classifiers [1]. It naturally provides class probability estimates [1] and can be easily extended to multiclass classification problems [2]. In this paper, we only focus on binary classification problems. Given a training dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, where $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \{\pm 1\}$ (d is feature dimension), the logistic regression model can be obtained through empirical risk minimization of the logistic loss function $l_{log}(z) = \log(1 + \exp(-z))$, where z is the margin variable and the base of logarithm is 2 [3], [4]. Fig. 1(a) shows $l_{log}(z)$ from which we learn that $l_{log}(z)$ is an unbounded and convex loss function. However, in practice, noise and class imbalance in the training dataset can deteriorate the performance of logistic regression [5], [6].

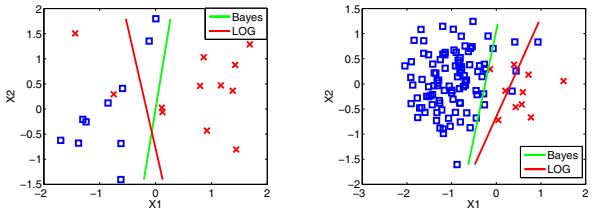
In some real-world datasets, the training samples are often contaminated by noise and some even have wrong labels. These are known as outliers which may arise due to instrument error, fraudulent behaviour and human error. Here, we treat outliers as the samples that locate in the other classes and are far away from their own classes because of feature noise or label noise [7], [8]. Since $l_{log}(z)$ is unbounded and outliers tend to have small margins, outliers usually have large losses and as a result, the decision boundary of logistic regression may deviate severely from the optimal one, which leads to the poor performance of logistic regression [9]. Fig. 2(a) intuitively shows that logistic regression is not robust to outliers since its decision boundary is biased because of the positive outliers. A number of methods have been proposed to tackle outliers and improve the robustness of classifiers. The straightforward methods are data preprocessing methods which aim to remove or relabel any suspect training samples [7], [10]. But these methods hold the risk of removing useful training samples and it is usually very hard to identify outliers in the high dimensional feature space. Assigning weights to the training samples is another popular approach to alleviate the detrimental effect of outliers [7], [11]. Many heuristic weight functions have been proposed which are used to show the importance of the training samples. Alternatively, the bounded and nonconvex loss functions are robust to outliers since they place bounded losses on outliers [7], [8], [9], [12], [13], [14]. In [9], the truncated logistic loss function $l_{tlog}(z) = \min\{l_{log}(z), l_{log}(q)\}$, where q specifies the truncation location, was proposed to improve the robustness of logistic regression to outliers. Moreover, Bootkrajang and Kaban [15] suggested to build the robust logistic regression by employing the latent variable model of label noise.

Class imbalance problems are also common in practical applications, e.g., medical diagnosis, credit fraud detection and oil spills [16], [17], [18]. In a class imbalance problem, the majority (negative) class has a large number of samples while the minority (positive) class only has a few. The positive class is usually of primary interest and suffers a greater misclassification cost than the negative class. Because logistic regression assumes balanced class distributions and misclassification costs, it tends to favour the negative class, resulting in a high classification accuracy but an unacceptably low detection rate of the positive class [6], which violates the goal of class imbalance problem. Fig. 2(b) intuitively shows that when classes are imbalanced, the decision boundary of logistic regression is biased to positives. A variety of class imbalance learning methods have been developed which can be broadly divided into data-level and algorithm-level methods [18]. Data-level methods are to balance the class distributions



(a) bounded logistic loss (b) C-loss

Fig. 1. Loss functions for classification.



(a) (b)

Fig. 2. Crosses are positives and squares are negatives. Positives and negatives are drawn from two Gaussian distributions with identity covariance matrix respectively. LOG is short for logistic regression. (a) Logistic regression is not robust to outliers. (b) Logistic regression is sensitive to class imbalance.

by data preprocessing, e.g., random undersampling, random oversampling and SMOTE [18], [19]. Algorithm-level methods focus on algorithmic modification so as to reduce their sensitiveness to class imbalance. Recognition-based learning [18] and cost-sensitive learning [11], [18], [20], [21] are two popular algorithm-level methods. Maalouf and Trafalis [6] proposed to deal with the class imbalance problem of logistic regression by assigning different error costs for different classes. Zhang and Hu [22] developed a new class imbalance learning framework called cost-free learning which is based on normalized mutual information.

Although there are many studies on outliers and class imbalance in isolation, only a few have addressed their combined effects [11], [20], [23], [24]. In [11], [24], heuristic weight functions are used to suppress the effect of outliers and different error costs are used to overcome the class imbalance problem. In this paper, we first propose a novel loss function called the bounded logistic loss function which is then used to develop a new robust logistic regression. Under the principle of cost-sensitive learning [17], different error costs are further used to overcome the sensitiveness of the new robust logistic regression to class imbalance [21], [24]. The main contribution of this work is as follows

1) Inspired by the Correntropy induced loss function (**C-loss**) [25], we develop the bounded logistic loss function which is a monotonic, bounded, smooth and nonconvex loss that is robust to outliers. Moreover, we show that $l_{log}(z)$ becomes a special case of the bounded logistic loss. Then, we construct a new robust logistic regression with the bounded logistic loss.
2) The new robust logistic regression is still sensitive to class imbalance. With cost-sensitive learning [18], [21], we reduce its sensitiveness to class imbalance by assigning different error costs for different classes. After combining the above two ideas, our proposed logistic regression model has the ability of handling both outliers and class imbalance.

Using the half-quadratic (**HQ**) optimization method [26], it is easy to optimize our proposed logistic regression model. And the experimental results confirm that it has better performance than the existing classifiers in the presence of both outliers and class imbalance. The paper is structured as follows. We develop the novel bounded logistic loss function in Section II and derive the proposed logistic regression model that is robust to outliers and class imbalance in Section III. Section IV presents the experimental results. Finally, we summarize the paper in Section V.

II. Bounded Logistic Loss Function

Correntropy is an information theoretic metric [27] and has been commonly used in robust learning [28]. Singh et al. [25] applied Correntropy to classification and developed C-loss

$$l_C(z) = \beta\Big(1 - \exp\big(-\frac{(1-z)^2}{2\sigma^2}\big)\Big), \tag{1}$$

where σ is the window width and $\beta = \Big(1-\exp\big(-\frac{1}{2\sigma^2}\big)\Big)^{-1}$ is a normalizing constant which ensures that $l_C(0) = 1$. $l_C(z)$ is a bounded, smooth and nonconvex loss, and embeds the higher order statistics of z. The least square (**LS**) loss function is [29]

$$l_{ls}(z) = (1-z)^2, \tag{2}$$

which is a popular convex loss function. Comparing $l_C(z)$ with $l_{ls}(z)$, we find that $l_C(z)$ can be rewritten as

$$l_C(z) = \beta\Big(1 - \exp\big(-\eta l_{ls}(z)\big)\Big), \tag{3}$$

where $\eta = \frac{1}{2\sigma^2} > 0$ is viewed as a scaling constant and $\beta = \frac{1}{1-\exp(-\eta)}$ is a normalizing constant. In other words, we can regard $l_C(z)$ as the bounded LS loss. Fig. 1(b) intuitively shows the relationship between $l_C(z)$ and $l_{ls}(z)$. According to this relationship, we derive the bounded logistic loss function

$$l_{blog}(z) = \beta\Big(1 - \exp\big(-\eta l_{log}(z)\big)\Big), \tag{4}$$

where $\eta > 0$ is also a scaling constant and $\beta = \frac{1}{1-\exp(-\eta)}$ is also a normalizing constant. Fig. 1(a) shows the different behaviours of $l_{log}(z)$ and $l_{blog}(z)$, and we can see that after transforming, the unbounded $l_{log}(z)$ becomes $l_{blog}(z)$ whose loss bound is determined by η. Furthermore, our $l_{blog}(z)$ is smooth while $l_{tlog}(z)$ is not. The relationship between $l_C(z)$ and $l_{ls}(z)$ is firstly used in [30] where the bounded hinge loss function is developed, which is then used to construct a new robust support vector machine. But there are two differences between [30] and this work: 1) [30] deals with support vector machine, while this work focuses on logistic regression; 2) class imbalance problem is not considered in [30], but is considered in this work. Proposition 1 states that $l_{blog}(z)$ becomes $l_{log}(z)$ as $\eta \to 0$.

Proposition 1. $\lim_{\eta\to 0} l_{blog}(z) = l_{log}(z)$.

Proof: According to the Taylor series of the exponential function, we know that

$$l_{blog}(z) = \sum_{j=1}^{+\infty} \frac{\beta\eta^j(-1)^{j+1}l_{log}^j(z)}{j!}.$$

According to L'Hopital's rule, we derive that

$$\lim_{\eta\to 0}\beta\eta^j = \lim_{\eta\to 0}\frac{\eta^j}{1-\exp(-\eta)} = \begin{cases} 1 & \text{if} \quad j = 1, \\ 0 & \text{if} \quad j \geq 2. \end{cases}$$

Consequently, $\lim_{\eta\to 0} l_{blog}(z) = l_{log}(z)$. ■

We show by Proposition 1 that $l_{log}(z)$ can be viewed as a special case of $l_{blog}(z)$. Because $l_{blog}(z)$ is bounded, the influence of outliers in each class can be effectively reduced. In the next section, using $l_{blog}(z)$, we derive the new robust logistic regression with different error costs, which can handle both outliers and class imbalance.

III. The Proposed Logistic Regression Model

We consider the following linear discriminant function

$$f(\boldsymbol{x}) = \boldsymbol{w}^T\phi(\boldsymbol{x}) + b, \tag{5}$$

where $\boldsymbol{w}$ is the weight vector; b is the bias term and $\phi(\boldsymbol{x})$ is the feature mapping function which is usually implicitly defined by a kernel function that satisfies the Mercer theory [31]. With $l_{blog}(z)$ and different error costs for different classes, the objective function considered here is

$$\min_{\boldsymbol{w},b} \frac{1}{2}(\boldsymbol{w}^T\boldsymbol{w} + b^2) + C^+ \sum_{\{i|y_i=+1\}}^{N} l_{blog}(z_i) + C^- \sum_{\{i|y_i=-1\}}^{N} l_{blog}(z_i), \tag{6}$$

where $z_i = y_i f(\boldsymbol{x}_i)$; C^+ and C^- are the error costs for the positive and negative respectively. We set $C^+ > C^-$ so as to reduce the effect of class imbalance [21]. By simple arithmetic modification, the above problem (6) is equivalent to the following problem

$$\begin{aligned}\max_{\boldsymbol{w},b} &-\frac{1}{2}(\boldsymbol{w}^T\boldsymbol{w} + b^2) + C^+\beta \sum_{\{i|y_i=+1\}}^{N} \exp\big(-\eta l_{log}(z_i)\big) \\ &+ C^-\beta \sum_{\{i|y_i=-1\}}^{N} \exp\big(-\eta l_{log}(z_i)\big).\end{aligned} \tag{7}$$

We recognize that the HQ optimization method [26] can be used to optimize (7), which has been successfully applied to robust face recognition [32] and robust feature extraction [33].

A. HQ Optimization for $l_{blog}(z)$

We consider a convex function $g(v) = -v\ln(-v) + v$, where $\ln(\cdot)$ denotes the natural logarithm and $v < 0$. According to the conjugate function theory [34], we have the following proposition.

Proposition 2. $\exp(-\eta l_{log}(z)) = \sup_{v<0}\{\eta l_{log}(z)v - g(v)\}$, where the supremum is achieved at $v = -\exp(-\eta l_{log}(z)) < 0$.

Proof: According to the definition of conjugate function [34], the conjugate function $g^*(u)$ of $g(v)$ is

$$g^*(u) = \sup_{v<0}\{uv - g(v)\} = \sup_{v<0}\{uv + v\ln(-v) - v\}.$$

Since $uv + v\ln(-v) - v$ is a concave function with respect to v, its supremum is achieved at $v = -\exp(-u) < 0$. Then, $(uv + v\ln(-v) - v)|_{v=-\exp(-u)} = \exp(-u)$. Consequently, we have

$$g^*(u) = \sup_{v<0}\{uv - g(v)\} = \exp(-u),$$

where the supremum is achieved at $v = -\exp(-u) < 0$. If we define $u = \eta l_{log}(z)$, then we have that

$$g^*(\eta l_{log}(z)) = \sup_{v<0}\{\eta l_{log}(z)v - g(v)\} = \exp(-\eta l_{log}(z)),$$

where the supremum is achieved at $v = -\exp(-\eta l_{log}(z)) < 0$. ■

With Proposition 2, we can derive the following equation

$$\begin{aligned}&-\frac{1}{2}(\boldsymbol{w}^T\boldsymbol{w} + b^2) + C^+\beta \sum_{\{i|y_i=+1\}}^{N} \exp\big(-\eta l_{log}(z_i)\big) \\ &+ C^-\beta \sum_{\{i|y_i=-1\}}^{N} \exp\big(-\eta l_{log}(z_i)\big) \\ =&-\frac{1}{2}(\boldsymbol{w}^T\boldsymbol{w} + b^2) + C^+\beta \sum_{\{i|y_i=+1\}}^{N} \sup_{v_i<0}\{\eta l_{log}(z_i)v_i - g(v_i)\} \\ &+ C^-\beta \sum_{\{i|y_i=-1\}}^{N} \sup_{v_i<0}\{\eta l_{log}(z_i)v_i - g(v_i)\} \\ =&-\frac{1}{2}(\boldsymbol{w}^T\boldsymbol{w} + b^2) + \sup_{\boldsymbol{v}\prec\boldsymbol{0}}\{C^+\beta \sum_{\{i|y_i=+1\}}^{N} \big(\eta l_{log}(z_i)v_i - g(v_i)\big) \\ &+ C^-\beta \sum_{\{i|y_i=-1\}}^{N} \big(\eta l_{log}(z_i)v_i - g(v_i)\big)\} \\ =&\sup_{\boldsymbol{v}\prec\boldsymbol{0}}\{-\frac{1}{2}(\boldsymbol{w}^T\boldsymbol{w} + b^2) + C^+\beta \sum_{\{i|y_i=+1\}}^{N} \big(\eta l_{log}(z_i)v_i - g(v_i)\big) \\ &+ C^-\beta \sum_{\{i|y_i=-1\}}^{N} \big(\eta l_{log}(z_i)v_i - g(v_i)\big)\},\end{aligned} \tag{8}$$

where $\boldsymbol{v} \in \mathbb{R}^N$. In (8), the second equation establishes since $\eta l_{log}(z_i)v_i - g(v_i), i = 1, 2, \cdots, N$ are independent functions in terms of v_i, and the third equation establishes since $-\frac{1}{2}(\boldsymbol{w}^T\boldsymbol{w}+b^2)$ is a constant with respect to v_i. Using (8), we can derive that (7) is equivalent to the following problem

$$\max_{\boldsymbol{w},b,\boldsymbol{v}\prec\boldsymbol{0}} R(\boldsymbol{w}, b, \boldsymbol{v}), \tag{9}$$

where

$$\begin{aligned}R(\boldsymbol{w}, b, \boldsymbol{v}) = &-\frac{1}{2}(\boldsymbol{w}^T\boldsymbol{w} + b^2) + C^+\beta \sum_{\{i|y_i=+1\}}^{N} \big(\eta l_{log}(z_i)v_i - g(v_i)\big) \\ &+ C^-\beta \sum_{\{i|y_i=-1\}}^{N} \big(\eta l_{log}(z_i)v_i - g(v_i)\big).\end{aligned}$$

After having (9), we can use the alternating optimization method. Specifically, given $(\boldsymbol{w}, b)$, we optimize over $\boldsymbol{v}$, and then given $\boldsymbol{v}$, we optimize over $(\boldsymbol{w}, b)$. First, suppose that we have $(\boldsymbol{w}^s, b^s)$ (the superscript s denotes the s-th iteration), then (9) is equivalent to

$$\begin{aligned}\max_{\boldsymbol{v}^s\prec\boldsymbol{0}} &C^+\beta \sum_{\{i|y_i=+1\}}^{N} \big(\eta l_{log}(z_i^s)v_i^s - g(v_i^s)\big) \\ &+ C^-\beta \sum_{\{i|y_i=-1\}}^{N} \big(\eta l_{log}(z_i^s)v_i^s - g(v_i^s)\big).\end{aligned} \tag{10}$$

According to Proposition 2, we know that the analytic solutions to (10) are

$$v_i^s = -\exp\big(-\eta l_{log}(z_i^s)\big) < 0,\ i = 1, 2, \cdots, N. \tag{11}$$

Second, after we have $\boldsymbol{v}^s$, we can obtain $(\boldsymbol{w}^{s+1}, b^{s+1})$ by solving the following problem[1]

$$\max_{\boldsymbol{w},b} -\frac{1}{2}(\boldsymbol{w}^T\boldsymbol{w}+b^2)+C^+\beta \sum_{\{i|y_i=+1\}}^{N} \eta l_{log}(z_i)v_i+C^-\beta \sum_{\{i|y_i=-1\}}^{N} \eta l_{log}(z_i)v_i, \tag{12}$$

[1]Here, we omit the superscripts of $(\boldsymbol{w}^{s+1}, b^{s+1})$, z_i^{s+1} and v_i^s for the reason of clarity.

Algorithm 1 The HQ optimization method for (6)

Input:

The training dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$; η in $l_{blog}(z)$; the error costs C^+ and C^-; the kernel function $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$; and the iteration number S.

Output: $(\boldsymbol{w}, b)$ or $(\boldsymbol{\alpha}, b)$.

1: Construct the Gram matrix K;
2: Set $s := 0$ and initialize $\boldsymbol{v}^s$;
3: **while** $s < S$ **do**
4: Obtain $(\boldsymbol{w}^{s+1}, b^{s+1})$ by solving (13) or $(\boldsymbol{\alpha}^{s+1}, b^{s+1})$ by (14);
5: Update $\boldsymbol{v}^{s+1}$ according to (11);
6: Set $s := s + 1$;
7: **end while**
8: **return** $(\boldsymbol{w}^s, b^s)$ or $(\boldsymbol{\alpha}^s, b^s)$.

which is equivalent to

$$\min_{\boldsymbol{w},b} \frac{1}{2}(\boldsymbol{w}^T\boldsymbol{w} + b^2) + \sum_{\{i|y_i=+1\}}^{N} C_i^+ l_{log}(z_i) + \sum_{\{i|y_i=-1\}}^{N} C_i^- l_{log}(z_i), \tag{13}$$

where $C_i^+ = C^+\beta\eta(-v_i) > 0$ and $C_i^- = C^-\beta\eta(-v_i) > 0$. We can see that (13) can be viewed as a weighted logistic regression problem. And $0 < -v_i < 1$ can show the relative importance of the training samples. In fact, (11) can also be considered as a weight function. For any training sample with large $l_{log}(z_i)$, $-v_i$ will be very small so as to reduce its detrimental influence. Since outliers usually have large $l_{log}(z_i)$'s, their $(-v_i)$'s are very small and then their influence is reduced. Compared with the existing heuristic weight functions [7], [11], our $(-v_i)$ is derived from $l_{blog}(z)$ which, we think, is more rational.

When using linear kernel, we suggest to apply LIBLINEAR [35] to help solve (13). When using nonlinear kernel, e.g., the Gaussian kernel, we usually have no idea about the explicit form of $\phi(\boldsymbol{x})$. Using the fact that $\boldsymbol{w} = \sum_{i=1}^N \alpha_i\phi(\boldsymbol{x}_i)$, (13) can be further transformed to the following problem [3]

$$\min_{\boldsymbol{\alpha},b} \frac{1}{2}(\boldsymbol{\alpha}^T K\boldsymbol{\alpha} + b^2) + \sum_{\{i|y_i=+1\}}^{N} C_i^+ l_{log}(z_i) + \sum_{\{i|y_i=-1\}}^{N} C_i^- l_{log}(z_i), \tag{14}$$

where K is the Gram matrix; $K_{ij} = \phi(\boldsymbol{x}_i)^T\phi(\boldsymbol{x}_j) = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, where $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is a kernel function; and $f(\boldsymbol{x}_i) = \sum_{j=1}^N \alpha_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) + b$. Quasi Newton method is applied to solve (14) [4].

Now, we derive the HQ optimization method for (6) and Algorithm 1 presents the whole optimization procedure. In Algorithm 1, $\boldsymbol{v}^0$ can be initialized by either setting it to $-\mathbf{1}$ or using some existing weight functions [7], [11].

B. Convergence of HQ Optimization

In this section, we present the convergence property of Algorithm 1. Proposition 3 states that Algorithm 1 converges.

Proposition 3. The sequence $\{R(\boldsymbol{w}^s, b^s, \boldsymbol{v}^s), s = 1, 2, \cdots, \}$ generated by Algorithm 1 converges.

Proof: According to (8), we derive that $R(\boldsymbol{w}, b, \boldsymbol{v}) \leq C^+\beta N^+ + C^-\beta N^-$, where N^+ is the number of positives and N^- is the number of negatives, which means that $R(\boldsymbol{w}, b, \boldsymbol{v})$ is upper bounded. Then, according to (10) and (12), we have that

$$R(\boldsymbol{w}^s, b^s, \boldsymbol{v}^s) \leq R(\boldsymbol{w}^{s+1}, b^{s+1}, \boldsymbol{v}^s) \leq R(\boldsymbol{w}^{s+1}, b^{s+1}, \boldsymbol{v}^{s+1}),$$

which shows that the sequence $\{R(\boldsymbol{w}^s, b^s, \boldsymbol{v}^s), s = 1, 2, \cdots, \}$ generated by Algorithm 1 is nondecreasing. Consequently, the sequence $\{R(\boldsymbol{w}^s, b^s, \boldsymbol{v}^s), s = 1, 2, \cdots, \}$ of Algorithm 1 converges. ■

There is a common problem with the existing heuristic weight functions which is that there is no guarantee for convergence [7], [11]. Hence, weights are usually updated only once. However, if we update weights according to (11), Proposition 3 ensures the convergence property.

We call the proposed logistic regression model derived in this section robust bounded logistic regression with different error costs (**RBLOG-DEC**).

TABLE I. CHARACTERISTICS OF DATASETS

	Dataset	# of features	# of samples	# of positives (%)
D1	Phoneme	5	5404	1586 (29.35)
D2	Vehicle(1 vs. others)	18	846	212 (25.06)
D3	Marketing (8 vs. others)	13	6876	1069 (15.55)
D4	OptDigits(0 vs. others)	64	5620	554 (9.86)
D5	Satimage(4 vs. others)	36	6435	626 (9.73)
D6	Vowel(0 vs. others)	10	990	90 (9.09)
D7	Letter(A vs. B $\sim$ N)	16	10723	789 (7.36)
D8	Mammography	6	11180	260 (2.33)

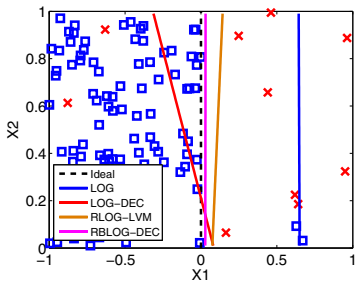


Fig. 3. Linear decision boundaries of different classifiers on a synthetic dataset.

IV. EXPERIMENTS

In this section, we carry out experiments to show the performance of RBLOG-DEC. The datasets used in our experiments are shown in Table I. Except the Mammography dataset, which is generously provided by Dr. Nitesh Chawla [19], the other datasets are from the UCI machine learning repository [36]. We learn from Table I that class distributions in these datasets are imbalanced. First, we would like to present a toy example in order to intuitively show the effectiveness of RBLOG-DEC in the presence of outliers and class imbalance.

A. A Toy Example

We generate a synthetic dataset which is shown in Fig. 3. The positives are uniformly distributed on the right plane; the

negatives are uniformly distributed on the left plane; and the black dash line between the two planes is the ideal decision boundary. However, in this synthetic dataset, the class ratio of positives over negatives is $1:10$. Besides, there are 2 positive outliers and 2 negative outliers. We run logistic regression (**LOG**) [4], logistic regression with different error costs (**LOG-DEC**) [6], robust logistic regression with latent variable model of label noise (**RLOG-LVM**) [15] and RBLOG-DEC on this synthetic dataset. Linear kernel is used and for RBLOG-DEC, $\eta = 1$, $C^+ = 10$, $C^- = 1$, $S = 4$ and $\boldsymbol{v}^0 = -\mathbf{1}$.

The linear decision boundaries of these classifiers are shown in Fig. 3. First, the decision boundary of LOG is seriously biased from the ideal one and misclassifies most of the positives. Second, LOG-DEC has the mechanism to tackle class imbalance but due to the outliers, its decision boundary still does not approximate the ideal one well and misclassifies many negatives. Third, RLOG-LVM is designed to deal with label noise and we can see from Fig. 3 that the outliers do not influence its decision boundary much. But its decision boundary is still not satisfactory since it is near the positives. Fourth, the decision boundary of our proposed RBLOG-DEC can approximate the ideal one very well, which proves the effectiveness of RBLOG-DEC. The $(-v_i)$'s of the 4 outliers at the last iteration (0.0030, 0.0132, 0.0153 and 0.0169) are much smaller than those of normal samples (greater than 0.3656). Hence, the detrimental influence of the 4 outliers is efficiently reduced. And the different error costs C^+ and C^- ensure that its decision boundary is not near the positives. To conclude, we can expect good performance of RBLOG-DEC on a dataset with outliers and class imbalance.

B. Comparison on the Real-World Datasets

In this section, we compare the performance of RBLOG-DEC with other classifiers on the real-world datasets in Table I. First, we list all the classifiers in comparison and their parameter settings so as to make the comparison clear.

1) Classifiers in Comparison: We compare the following classifiers

- LOG [3], [4].
- LOG-DEC [6]. The negative error cost is fixed to be 1 and we modify the positive error cost.
- The support vector machine with different error costs (**SVM-DEC**) [21]. Like LOG-DEC, the negative error cost is set to be 1 and we vary the positive error cost.
- RLOG-LVM [15].
- The classifiers based on the asymmetric stagewise least square loss function (**ASLS**) [20]. The negative ramp and margin coefficients are both set to be 1 and we change the positive ramp and margin coefficients.
- Our proposed RBLOG-DEC. We set $S = 4$ by balancing the computational complexity and the convergence of HQ optimization method. $C^- = 1$ and

$$-v_i^0 = \frac{2}{1 + \exp(\theta \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|_2)},$$

where $\theta > 0$ is a given constant and $\bar{\boldsymbol{x}}$ is the class center of the class that $\boldsymbol{x}_i$ belongs to. We modify η and C^+ during the experiments.

LOG is selected to show that it does not perform well on the dataset with outliers and class imbalance. LOG-DEC and SVM-DEC are only designed to overcome class imbalance and we want to check their performance if outliers and class imbalance both occur. RLOG-LVM is only proposed to tackle outliers and we are eager to know its performance when class imbalance exists. ASLS is a classifier that can handle outliers and class imbalance at the same time, and we would like to compare it with our proposed RBLOG-DEC.

2) Experiment Design: The datasets in Table I are imbalanced but they generally have no outliers. In order to study the combined influence of outliers and class imbalance, we artificially introduce label noise into the training dataset [7], [8], [9], [12], [15], [23], [24]. Label noise is a kind of outliers which is the process of polluting labels [7]. Like the way of injecting label noise in [23], we flip $\rho\% \times N^+$ (N^+ is the number of positives) labels of positive to be negative, and also $\rho\% \times N^+$ labels of negative to be positive. We consider 2 levels of label noise, i.e., $\rho = 5$ and $\rho = 15$. The hyper-parameters of the above classifiers, including η in RBLOG-DEC, are tuned by 5-fold cross-validation on the training dataset. To evaluate classification results, we use *G-mean* and *F-measure* which are commonly used in the class imbalance problem [17], [18]

$$\text{G-mean} = \sqrt{\text{TPR} \times \text{TNR}},$$
$$\text{F-measure} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}},$$

where TPR stands for true positive rate and TNR stands for true negative rate.

Table II shows the 10-fold cross-validation results of all the classifiers with linear kernel and Table III shows the 10-fold cross-validation results of all the classifiers with the Gaussian kernel $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\gamma \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2)$. The Friedman test with Nemenyi post-hoc test [37] is also used for statistical analysis. It is a nonparametric test which is commonly used to compare the performance of multiple classifiers over multiple datasets. Fig. 4 presents the critical difference (**CD**) diagrams of the Friedman test with Nemenyi post-hoc test at the 5% significance level. In Fig. 4, the lower the rank is, the better the classifier is. And if the difference between the ranks of two classifiers is greater than CD, their performance is considered to be significantly different. In Fig. 4, groups of classifiers that are not significantly different are connected by a thick line.

3) LOG vs. RBLOG-DEC: LOG usually has good performance on the clean and balanced datasets, but when there are outliers and class imbalance, its performance is not satisfactory, which is confirmed by the results in Table II and Table III. In Table II, LOG with linear kernel performs very badly on the Marketing and Satimage datasets, and in Table III, LOG with the Gaussian kernel performs very badly on the Marketing dataset. However, we learn from Fig. 4 that our proposed classifier, RBLOG-DEC, can perform statistically better than LOG in all the cases, which demonstrates that compared with LOG, RBLOG-DEC is effective in handling outliers and class imbalance.

4) LOG-DEC and SVM-DEC vs. RBLOG-DEC: LOG-DEC [6] and SVM-DEC [21] are originally designed to deal with class imbalance problem and here, we check their

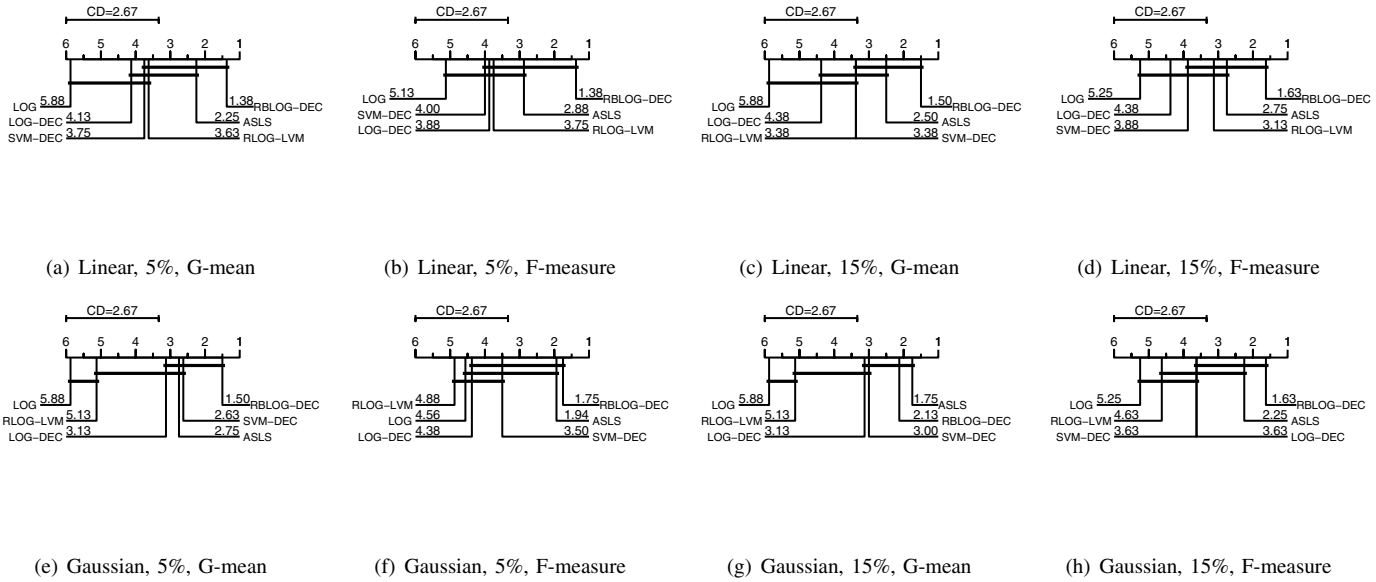
TABLE II. THE 10-FOLD CROSS-VALIDATION RESULTS OF ALL THE CLASSIFIERS WITH LINEAR KERNEL. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD. (MEAN(STD)%)

(a) 5% label noise level

Dataset		LOG	LOG-DEC	SVM-DEC	RLOG-LVM	ASLS	RBLOG-DEC
Phoneme	G-mean	61.19(4.77)	74.09(2.71)	74.33(1.84)	75.50(2.60)	75.61(3.29)	**76.08(2.70)**
	F-measure	49.56(6.10)	62.65(3.16)	63.02(2.17)	64.33(3.05)	65.01(4.02)	**65.02(3.16)**
Vehicle	G-mean	61.68(6.24)	77.80(3.70)	78.88(3.57)	77.18(3.40)	79.04(3.59)	**79.84(3.32)**
	F-measure	49.99(7.76)	63.33(4.39)	64.45(4.52)	62.38(4.17)	**66.06(5.28)**	65.61(4.60)
Marketing	G-mean	1.93(4.08)	69.05(2.46)	67.44(2.80)	69.06(2.08)	68.74(2.16)	**69.47(2.29)**
	F-measure	0.37(0.78)	40.41(2.46)	38.73(2.65)	40.81(2.28)	40.79(3.30)	**40.89(2.34)**
OptDigits	G-mean	98.42(1.22)	98.19(0.75)	98.91(0.93)	98.66(1.19)	99.02(1.08)	**99.33(0.67)**
	F-measure	98.16(1.36)	98.16(1.36)	98.45(1.22)	95.94(1.81)	98.64(1.31)	**98.82(1.05)**
Satimage	G-mean	9.02(8.29)	70.43(1.89)	69.15(1.60)	**70.67(2.21)**	70.24(1.72)	70.54(2.12)
	F-measure	2.76(3.01)	29.78(1.52)	29.07(1.03)	**30.32(1.71)**	29.66(1.65)	29.82(1.58)
Vowel	G-mean	79.00(7.44)	92.50(3.18)	92.61(2.79)	92.30(4.64)	**93.24(2.95)**	93.04(4.49)
	F-measure	74.09(9.37)	77.33(8.29)	75.15(8.03)	70.36(8.44)	75.98(8.77)	**78.61(9.68)**
Letter	G-mean	91.21(2.58)	95.29(0.91)	95.52(1.24)	95.30(1.08)	**95.56(0.94)**	95.55(1.09)
	F-measure	88.18(2.74)	88.18(2.74)	89.22(2.69)	86.09(3.18)	88.92(2.84)	**91.16(2.78)**
Mammography	G-mean	56.59(8.26)	86.84(2.86)	86.53(3.55)	78.72(7.71)	87.59(2.87)	**88.14(4.18)**
	F-measure	46.55(10.18)	45.73(11.34)	40.88(3.10)	**56.52(8.98)**	41.59(2.86)	54.83(9.80)

(b) 15% label noise level

Dataset		LOG	LOG-DEC	SVM-DEC	RLOG-LVM	ASLS	RBLOG-DEC
Phoneme	G-mean	54.49(5.16)	74.21(2.91)	73.60(1.77)	75.93(2.64)	75.64(3.38)	**75.95(2.28)**
	F-measure	42.01(6.48)	62.81(3.40)	62.34(1.96)	64.87(3.12)	65.05(4.09)	**65.07(2.67)**
Vehicle	G-mean	55.23(11.91)	76.04(5.32)	77.38(6.13)	75.99(3.91)	76.17(4.66)	**78.10(5.26)**
	F-measure	43.17(14.46)	61.88(7.12)	63.22(7.57)	62.16(5.42)	63.01(6.01)	**64.00(6.48)**
Marketing	G-mean	0.00(0.00)	68.45(1.84)	66.32(2.76)	**69.35(2.63)**	68.58(2.03)	68.78(2.13)
	F-measure	0.00(0.00)	39.74(1.88)	37.68(2.61)	**41.98(2.97)**	41.13(3.12)	41.09(2.18)
OptDigits	G-mean	97.42(1.39)	97.15(1.14)	**98.83(0.50)**	97.58(1.74)	98.64(0.51)	98.72(1.08)
	F-measure	97.22(1.55)	97.22(1.55)	98.35(1.14)	93.70(2.74)	**98.45(0.97)**	98.35(0.84)
Satimage	G-mean	3.80(6.12)	68.88(2.30)	66.70(1.35)	**71.17(1.80)**	69.63(1.67)	70.46(2.15)
	F-measure	0.94(1.52)	28.64(1.68)	27.78(0.78)	**30.57(1.39)**	29.51(1.37)	29.75(1.57)
Vowel	G-mean	62.91(12.50)	90.13(2.66)	90.41(4.33)	89.98(6.25)	**92.72(2.88)**	91.26(4.93)
	F-measure	55.68(15.99)	54.51(3.72)	56.59(5.81)	61.31(7.00)	64.83(4.31)	**77.67(10.03)**
Letter	G-mean	88.56(3.54)	94.59(0.95)	95.20(1.12)	95.08(1.20)	95.25(1.07)	**95.57(1.12)**
	F-measure	86.68(3.60)	86.68(3.60)	87.74(3.06)	84.58(5.41)	87.72(2.95)	**89.99(3.11)**
Mammography	G-mean	53.24(9.76)	85.81(3.51)	86.05(3.29)	80.77(5.87)	86.93(3.55)	**87.94(3.97)**
	F-measure	42.79(11.87)	43.30(11.42)	40.05(2.96)	**45.59(7.85)**	39.54(3.35)	43.72(11.08)



(a) Linear, 5%, G-mean (b) Linear, 5%, F-measure (c) Linear, 15%, G-mean (d) Linear, 15%, F-measure

(e) Gaussian, 5%, G-mean (f) Gaussian, 5%, F-measure (g) Gaussian, 15%, G-mean (h) Gaussian, 15%, F-measure

Fig. 4. The CD diagrams of the Friedman test with Nemenyi post-hoc test at the 5% significance level. Groups of classifiers that are not significantly different are connected. In the above subtitles, 5% and 15% represent label noise levels.

TABLE III. THE 10-FOLD CROSS-VALIDATION RESULTS OF ALL THE CLASSIFIERS WITH THE GAUSSIAN KERNEL. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD. (MEAN(STD)%)

(a) 5% label noise level

Dataset		LOG	LOG-DEC	SVM-DEC	RLOG-LVM	ASLS	RBLOG-DEC
Phoneme	G-mean	81.42(1.73)	84.90(2.11)	84.98(2.16)	80.25(2.11)	85.49(2.26)	**85.67(2.20)**
	F-measure	75.11(1.88)	75.84(2.48)	75.74(2.58)	73.59(2.20)	**77.49(2.60)**	76.99(2.63)
Vehicle	G-mean	62.26(5.09)	79.92(3.39)	80.78(3.80)	65.10(6.27)	81.10(4.48)	**81.45(3.64)**
	F-measure	51.20(6.39)	66.29(3.81)	67.07(5.00)	53.54(7.63)	67.87(5.56)	**67.98(4.75)**
Marketing	G-mean	0.97(3.07)	70.13(3.21)	69.93(3.29)	62.58(6.08)	69.60(3.38)	**70.56(3.36)**
	F-measure	0.19(0.59)	41.99(2.96)	41.32(3.21)	39.61(4.92)	41.36(3.65)	**42.30(3.50)**
OptDigits	G-mean	99.28(0.72)	**99.75(0.40)**	99.66(0.44)	99.37(0.85)	99.65(0.44)	99.67(0.63)
	F-measure	99.27(0.72)	99.20(0.78)	99.28(0.57)	98.74(0.88)	99.37(0.61)	**99.46(0.64)**
Satimage	G-mean	69.76(2.82)	89.20(1.56)	89.35(1.54)	70.03(4.04)	89.36(1.65)	**89.89(1.08)**
	F-measure	59.33(4.03)	59.19(2.46)	60.70(2.73)	58.79(5.74)	**62.61(3.01)**	61.07(2.06)
Vowel	G-mean	96.49(4.96)	99.22(0.76)	**99.61(0.59)**	98.19(2.23)	99.44(0.79)	99.27(0.75)
	F-measure	96.32(5.23)	95.80(5.09)	96.52(5.15)	93.25(4.18)	96.32(5.23)	**96.66(5.92)**
Letter	G-mean	97.35(1.29)	99.53(0.46)	**99.78(0.28)**	98.42(1.09)	99.66(0.33)	99.69(0.23)
	F-measure	97.19(1.35)	96.10(1.26)	98.75(0.88)	97.95(1.52)	**99.43(0.36)**	99.00(1.22)
Mammography	G-mean	69.52(6.70)	90.77(3.85)	89.67(4.11)	74.72(4.04)	90.06(4.09)	**90.86(3.71)**
	F-measure	61.26(8.38)	58.89(8.15)	58.79(8.91)	**66.10(5.08)**	63.82(6.94)	61.13(9.70)

(b) 15% label noise level

Dataset		LOG	LOG-DEC	SVM-DEC	RLOG-LVM	ASLS	RBLOG-DEC
Phoneme	G-mean	78.83(2.16)	83.98(2.02)	84.50(1.54)	77.18(2.97)	**85.00(1.79)**	84.05(1.62)
	F-measure	72.51(2.43)	74.58(2.35)	75.15(1.88)	70.91(3.39)	**76.24(2.30)**	76.18(1.89)
Vehicle	G-mean	52.90(7.20)	**78.61(2.71)**	78.16(3.73)	67.23(8.92)	78.59(4.19)	78.21(4.62)
	F-measure	41.30(9.31)	65.82(3.85)	64.64(5.80)	54.71(8.78)	64.12(7.13)	**65.86(5.99)**
Marketing	G-mean	0.00(0.00)	69.87(3.35)	70.12(3.00)	50.61(6.15)	68.66(3.82)	**70.27(3.24)**
	F-measure	0.00(0.00)	41.65(3.47)	41.58(3.06)	32.04(5.94)	40.67(3.74)	**42.02(3.46)**
OptDigits	G-mean	97.90(1.43)	99.42(0.59)	99.29(0.59)	98.57(0.97)	**99.50(0.44)**	99.44(0.43)
	F-measure	97.86(1.46)	98.05(1.34)	97.28(1.46)	97.31(1.37)	**99.08(0.97)**	98.87(0.73)
Satimage	G-mean	62.44(4.07)	88.54(1.76)	88.13(1.80)	64.01(3.83)	**89.00(1.54)**	88.70(1.66)
	F-measure	52.41(5.39)	56.79(2.41)	58.24(3.42)	53.75(5.55)	**60.58(3.59)**	60.08(2.41)
Vowel	G-mean	89.89(5.84)	97.40(1.15)	97.17(1.47)	95.66(5.16)	**98.76(1.46)**	97.82(1.15)
	F-measure	89.24(6.41)	89.42(6.14)	90.96(5.11)	85.44(7.85)	90.65(7.32)	**91.24(6.33)**
Letter	G-mean	94.24(1.56)	98.88(0.60)	**99.61(0.32)**	97.64(1.27)	99.38(0.36)	99.36(0.39)
	F-measure	94.07(1.65)	94.07(1.65)	97.41(1.30)	97.15(1.75)	97.77(1.59)	**98.17(1.84)**
Mammography	G-mean	66.12(3.86)	89.77(4.95)	89.81(4.39)	71.33(3.81)	90.16(4.76)	**91.20(4.03)**
	F-measure	57.76(9.26)	57.76(9.26)	57.17(8.09)	**62.80(5.09)**	59.93(7.62)	59.69(13.17)

performance on the datasets with both outliers and class imbalance. According to Fig. 4, both LOG-DEC and SVM-DEC have better performance than LOG, which proves that they can handle class imbalance problem. However, since the loss functions used in LOG-DEC and SVM-DEC are not robust to outliers, they are sensitive to outliers and this is the reason that they do not perform as well as RBLOG-DEC in our experiments. Fig. 4 indicates that SVM-DEC generally performs slightly better than LOG-DEC.

5) RLOG-LVM vs. RBLOG-DEC: RLOG-LVM [15] uses a latent variable model of label noise to reduce the influence of outliers. Classification results in Table II and Table III confirm that after using the latent variable model of label noise, RLOG-LVM generally achieves better performance than LOG. Although RLOG-LVM may perform better than RBLOG-DEC on some datasets, like the Satimage dataset in Table II, we know from Fig. 4 that RBLOG-DEC is better overall. The advantage of RBLOG-DEC over RLOG-LVM is that RBLOG-DEC has another mechanism to deal with class imbalance while RLOG-LVM does not.

6) ASLS vs. RBLOG-DEC: ASLS approaches the squared ramp loss by updating targets of samples. In ASLS, there are two coefficients, namely ramp and margin coefficients. The ramp coefficient determines the loss bound and the margin coefficient is in charge of the distances to decision boundary of different classes. By setting different ramp and margin coefficients for different classes, ASLS has the ability of tackling outliers and class imbalance. Fig. 4 suggests that ASLS generally has better performance than LOG, LOG-DEC, SVM-DEC and RLOG-LVM, which confirms its effectiveness in the case of outliers and class imbalance. But compared with our RBLOG-DEC, its performance is still slightly worse, which, we think, is related with the heuristic idea of ASLS in some degree.

To summarize, a classifier should have both mechanisms to deal with outliers and class imbalance in order to achieve good performance when there are outliers and class imbalance in the datasets. Our proposed RBLOG-DEC uses the bounded logistic loss to suppress the influence of outliers and different error costs for different classes to solve the class imbalance problem. The above experimental results demonstrate that compared with the existing classifiers, RBLOG-DEC performs better on the dataset with both outliers and class imbalance.

V. CONCLUSION

Outliers and class imbalance are two common problems in practical applications. In this paper, we propose to deal with the problems of logistic regression with outliers and class imbalance. First, we develop the bounded logistic loss $l_{blog}(z)$ by employing the relationship between $l_C(z)$ and $l_{ls}(z)$. The

bounded logistic loss $l_{blog}(z)$ is a monotonic, bounded and smooth loss that is robust to outliers, which is then used to develop a new robust logistic regression. Second, with the principle of cost-sensitive learning, different error costs for different classes is further used to deal with class imbalance problem. By combining the above two ideas, we propose the robust bounded logistic regression with different error costs which is RBLOG-DEC for short. According to the experimental results on the real-world datasets with different degrees of outliers and class imbalance, RBLOG-DEC is demonstrated to have better performance than the existing classifiers.

In the future, we will explore the usage of the bounded hinge loss [30], [38] and bounded exponential loss in the class imbalance problem, and then compare their performance with RBLOG-DEC proposed in this paper.

ACKNOWLEDGMENT

This work is supported in part by NSFC grants # 61273196 and # 61573348.

REFERENCES

[1] C. Perlich, F. Provost, and J. S. Simonoff, "Tree induction vs. logistic regression: a learning-curve analysis," *The Journal of Machine Learning Research*, vol. 4, pp. 211–255, Dec. 2003.

[2] B. Krishnapuram, L. Carin, M. A. Figueiredo, and A. J. Hartemink, "Sparse multinomial logistic regression: fast algorithms and generalization bounds," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 957 – 968, Jun. 2005.

[3] S. S. Keerthi, K. B. Duan, S. K. Shevade, and A. N. Poo, "A fast dual algorithm for kernel logistic regression," *Machine Learning*, vol. 61, no. 1, pp. 151–165, Jul. 2005.

[4] C.-J. Lin, R. C. Weng, and S. S. Keerthi, "Trust region newton method for logistic regression," *The Journal of Machine Learning Research*, vol. 9, pp. 627–650, Jun. 2008.

[5] D. Pregibon, "Logistic regression diagnostics," *The Annals of Statistics*, vol. 9, no. 4, pp. 705–724, Jul. 1981.

[6] M. Maalouf and T. B. Trafalis, "Robust weighted kernel logistic regression in imbalanced and rare events data," *Computational Statistics & Data Analysis*, vol. 55, no. 1, pp. 168–183, Jan. 2011.

[7] B. Frenay and M. Verleysen, "Classification in the presence of label noise: a survey," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 25, no. 5, pp. 845–869, May 2014.

[8] Y. C. Wu and Y. F. Liu, "Robust truncated hinge loss support vector machines," *Journal of the American Statistical Association*, vol. 102, no. 479, pp. 974–983, Sep. 2007.

[9] S. Y. Park and Y. F. Liu, "Robust penalized logistic regression with truncated loss functions," *Canadian Journal of Statistics*, vol. 39, no. 2, pp. 300–323, Jun. 2011.

[10] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: a survey," *ACM Computing Surveys*, vol. 41, no. 3, Jul. 2009.

[11] R. Batuwita and V. Palade, "Fsvm-cil: fuzzy support vector machines for class imbalance learning," *IEEE Trans. on Fuzzy Systems*, vol. 18, no. 3, pp. 558–571, Jun. 2010.

[12] H. Masnadi-Shirazi and N. Vasconcelos, "On the design of loss functions for classification: theory, robustness to outliers, and savageboost," in *Proc. of Advances in Neural Information Processing Systems (NIPS)*, 2008, pp. 1049–1056.

[13] S.-H. Yang and B.-G. Hu, "A stagewise least square loss function for classification." in *Proc. of SIAM International Conference on Data Mining (SDM)*, 2008, pp. 120–131.

[14] Q. G. Miao, Y. Cao, G. Xia, M. G. Gong, J. C. Liu, and J. F. Song, "Rboost: label noise-robust boosting algorithm based on a nonconvex loss function and the numerically stable base learners," *IEEE Trans. on Neural Networks and Learning Systems*, vol. PP, no. 99, Sep. 2015.

[15] J. Bootkrajang and A. Kabán, "Learning kernel logistic regression in the presence of class label noise," *Pattern Recognition*, vol. 47, no. 11, pp. 3641–3655, Nov. 2014.

[16] S. Wang, L. L. Minku, and X. Yao, "Resampling-based ensemble methods for online class imbalance learning," *IEEE Trans. on Knowledge and Data Engineering*, vol. 27, no. 5, pp. 1356–1368, May 2015.

[17] H. B. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[18] H. B. He and Y. Q. Ma, *Imbalanced Learning: Foundations, Algorithms, and Applications*. Wiley-IEEE Press, 2013.

[19] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002.

[20] G. B. Xu, B.-G. Hu, and J. C. Principe, "An asymmetric stagewise least square loss function for imbalanced classification," in *Proc. International Joint Conference on Neural Networks (IJCNN)*, 2014, pp. 1107–1114.

[21] K. Veropoulos, C. Campbell, and N. Cristianini, "Controlling the sensitivity of support vector machines," in *Proc. of International Joint Conference on Artificial Intelligence (IJCAI)*, 1999, pp. 55–60.

[22] X. W. Zhang and B.-G. Hu, "A new strategy of cost-free learning in the class imbalance problem," *IEEE Trans. on Knowledge and Data Engineering*, vol. 26, no. 12, pp. 2872–2885, Dec. 2014.

[23] T. M. Khoshgoftaar, J. V. Hulse, and A. Napolitano, "Comparing boosting and bagging techniques with noisy and imbalanced data," *IEEE Trans. on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 41, no. 3, pp. 552–568, May 2011.

[24] S. R. Pan and X. Q. Zhu, "Graph classification with imbalanced class distributions and noise," in *Proc. of International Joint Conference on Artificial Intelligence (IJCAI)*, 2013, pp. 1586–1592.

[25] A. Singh, R. Pokharel, and J. C. Principe, "The c-loss function for pattern classification," *Pattern Recognition*, vol. 47, no. 1, pp. 441–453, Jan. 2014.

[26] M. Nikolova and M. K. Ng, "Analysis of half-quadratic minimization methods for signal and image recovery," *SIAM Journal on Scientific Computing*, vol. 27, no. 3, pp. 937–966, Mar. 2005.

[27] W. Liu, P. P. Pokharel, and J. C. Principe, "Correntropy: Properties and applications in non-gaussian signal processing," *IEEE Trans. on Signal Processing*, vol. 55, no. 11, pp. 5286–5298, Nov. 2007.

[28] J. C. Principe, *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*. Springer New York, 2010.

[29] T. V. Gestel, J. A. Suykens, B. Baesens, S. Viaene, J. Vanthienen, G. Dedene, B. D. Moor, and J. Vandewalle, "Benchmarking least squares support vector machine classifiers," *Machine Learning*, vol. 54, no. 1, pp. 5–32, Jan. 2004.

[30] G. B. Xu, Z. Cao, B.-G. Hu, and J. C. Principe, "Robust support vector machines based on the rescaled hinge loss function," *submitted to a journal*, 2015.

[31] V. Vapnik, *Statistical Learning Theory*. Wiley, 1998.

[32] R. He, W. S. Zheng, and B.-G. Hu, "Maximum correntropy criterion for robust face recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1561–1576, Aug. 2011.

[33] X.-T. Yuan and B.-G. Hu, "Robust feature extraction via information theoretic learning," in *Proc. of International Conference on Machine Learning (ICML)*, 2009, pp. 1193–1200.

[34] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[35] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: a library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, Jun. 2008.

[36] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[37] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, Dec. 2006.

[38] H. Masnadi-Shirazi and N. Vasconcelos, "Risk minimization, probability elicitation, and cost-sensitive svms," in *Proc. of International Conference on Machine Learning (ICML)*, 2010, pp. 759–766.