A Robust Loss Function for Multiclass Classification

Lei Zhao

***Abstract*—The application of robust loss function is an important approach to classify data sets that contaminated by noisy data points, in particular by outliers. In this paper we present an extension of smoothed 0-1 loss function to the multiclass case. In multiclass case, Fisher consistency of smoothed 0-1 loss function is satisfied. A classification algorithm is developed for multiclass classification problems. The performance of Hinge loss function and smoothed 0-1 loss function based classification algorithms are compared on several data sets with different levels of noise. Experiments show that smoothed 0-1 loss function demonstrates improved performance for data classification on more noisy datasets with noisy features or labels.**

***Index Terms*—Optimization, classification, loss function, robust.**

I. Introduction

Loss function plays an important role in data classification. In order to deal with data sets that contaminated by outliers, many robust loss functions including Hinge Loss Function[1], Exponential Loss Function [2], [3], Log Loss Function [3], [4], Mada boost Loss Function [5], Sigmoid Loss Function [6], Φ Loss Function [7], Ramp Loss Function [8], [9](also known as Robust Truncated Hinge Loss Function [10]), Normalized Sigmoid Lost Function [11], Logistic Difference Loss Function [12], have been proposed and evaluated. Based on the analysis of these loss functions, Smoothed 0-1 Loss Function has been proposed, applied and evaluated in binary data classification [13]. However, in many real applications we are more interested in solving multiclass problems where there are more than two classes.

This paper examines the application of Smoothed 0-1 Loss Function to multiclass classification problems. Fisher consistency of the smoothed 0-1 loss function is investigated. Based on the smoothed 0-1 loss function, a multiclass classification algorithm is proposed. We hypothesize that this algorithm will perform well on noisy data sets, in particular for those noisy data sets with many outliers. To prove this, we compare this algorithm with hinge loss based multiclass classification algorithm through a systematic evaluation on several well-known multiclass data sets corrupted with two kinds of noise: class noise and attribute noise. Our experiments indicate that the smoothed 0-1 loss function based multiclass classification algorithm is robust for label noise and attribute noise.

The structure of this paper is as follows. We first review some important loss functions for multiclass classification problem in Section II. Then Fisher consistent loss functions for multiclass classification is investigated in Section III. In Section IV, a single machine approach for multiclass data classification is proposed. Numerical experiments are conducted in Section V, followed by conclusions in Section VI.

Manuscript received September 17, 2013; revised November 1, 2013.

Lei Zhao is with the Office of Research, University of the Sunshine Coast, Maroochydore QLD 4558 Australia (e-mail: lzhao@usc.edu.au).

II. Multiclass Classification

Given a multiclass data set:

$$\begin{pmatrix} x_{11}x_{12} & \cdots & x_{1n}\, y_1 \\ \vdots & \ddots & \vdots \\ x_{l1}\, x_{l2} & \cdots & x_{ln} y_l \end{pmatrix}$$

where $x_{ij} \in R, y_i \in \{1, \dots, k\}$ denotes the class of the ith data point $x_i = (x_{i1}, \dots, x_{in})$,we consider the following linear multiclass data classification problem:

$$XW + B = F$$

where

$$X = \begin{pmatrix} x_{11}x_{12} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{l1}\, x_{l2} & \cdots & x_{ln}\} \end{pmatrix},$$

$$W = \begin{pmatrix} w_{11}w_{12} & \cdots & w_{1k} \\ \vdots & \ddots & \vdots \\ w_{n1}\, w_{n2} & \cdots & w_{nk} \end{pmatrix},$$

$$B = \begin{pmatrix} b_1 b_2 & \cdots & b_k \\ \vdots & \ddots & \vdots \\ b_1\, b_2 & \cdots & b_k \end{pmatrix},$$

$$F = \begin{pmatrix} f_{11}f_{12} & \cdots & f_{1k} \\ \vdots & \ddots & \vdots \\ f_{l1}f_{l2} & \cdots & f_{lk} \end{pmatrix},$$

And the combination of matrix W and vector $b_1, \dots, b_k$ is called linear classifier, which means given W and $b_1, \dots, b_k$ we can classify any data point $x_i = (x_{i1}, \dots, x_{in})$.

We define the following loss functions with respect to data point $\boldsymbol{x_i}$ and the corresponding class label y_i:

$$\text{Square Loss } V_{mc}(x_i, y_i) = \sum_{j \neq y_i} \left(f_{iy_i} - f_{ij} - 1\right)^2 \quad (1)$$

$$\text{Hinge Loss } V_{mc}(x_i, y_i) = \sum_{j \neq y_i} \left(1 - \left(f_{iy_i} - f_{ij}\right)\right)_+^2 \quad (2)$$

$$\text{Smoothed 0-1 Loss } V_{mc}(x_i, y_i) = \sum_{j \neq y_i} V\left(f_{iy_i} - f_{ij}\right)^2 \quad (3)$$

where loss function $V(\cdot)$ is defined as

$$V(t_j) = \begin{cases} 0, & t_j > 1 \\ \frac{1}{4}t_j^3 - \frac{3}{4}t\_j + \frac{1}{2}, & -1 \le t_j \le 1 \\ 1, & t_j < -1 \end{cases} \quad (4)$$

Given a single data point x that belongs to class i, for each

pair $i \neq j$ we pay a penalty which is measured by loss function$V(\cdot)$. This penalty plays an important role in the optimizationmodule, which will be further discussed in Section IV.

III. Fisher Consistent Loss Functions for Multiclass Classification

Fisher consistency plays a fundamental role in the construction of successful binary margin-based classifiers. It is possible to first design a nice Fisher consistent loss function and then construct the corresponding classifier [14]. A Fisher consistent smoothed 0-1 loss function for binary classification, constructed the corresponding classification algorithms and the experimental results show some good properties of the new algorithm [13]. Following the same line, we will extend this work to multiclass classification problems.

There has been a considerable amount of work in the literature to extend the binary classifiers to the multiclass case. In particular, a wide class of smooth convex loss functions that are Fisher consistency for multiclass classification has been investigated by Zou *et al.* [15].

A loss function $V(\cdot)$ is said to be Fisher consistency fork-class classification $(k \geq 3)\ if\ \forall x$ in a set of full measure, the following optimization problem

$$f^*(x) = \operatorname{argmin}_{f(x)} V\big(p, f(x)\big) \quad s.t. \sum_{j=1}^{k} f_j(x) = 0 \quad (5)$$

has a unique solution f^*, and

$$\operatorname*{argmin}_j f_j^*(x) = \operatorname*{argmin}_j p(y = j|x) \quad (6)$$

where

$$V(p, f(x)) = \sum_{j=1}^{k} V\left(f_j(x)\right) p(y = j|x) \quad (7)$$

is the expected V risk at x.

Remark 1: Unlike the margin $y_i f(x_i)$ defined in the binarycase, it is difficult to construct a special coding scheme for$y$ in multiclass classification. Zou *et al.* [15] proposed themulticlass margin as$f_{yi}(x_i)$. Thus the empirical risk is definedas that in (7).

Zou *et al.* [15] characterize a family of convex loss functions, including exponential loss and logistic regression loss(also called Logit loss), that are Fisher consistent for multiclass classification. Based on these multiclass Fisher consistent loss functions, Zou *et al.* [15] also derive some new multiclass boosting algorithms by minimizing the empirical loss. The results show that the corresponding algorithms converge tothe Bayes classifier in terms of classification error.

Theroem 1: [15] Let $V(\cdot)$ be a twice differentiable lossfunction, if $V'(0) < 0\ and\ V''(t) > 0$, then V is Fisher consistent. Moreover, letting f^* be the solution of (5), then we have

$$p(y = j|x) = \frac{1/V'(f_j^*(x))}{\sum_{l=1}^{k} 1/V'(f_l^*(x))} \quad (8)$$

Next we extend the work of Zou *et al.* [15] by considering the nonconvex loss function for multiclass classification problem.

Theroem 2: Let $V(\cdot)$ be a bounded continuous decreasingfunction. Then for any given x the following optimizationproblem

$$f^*(x) = \operatorname{argmin}_{f(x)} V\big(p, f(x)\big) \quad s.t. \sum_{j=1}^{k} f_j(x) = 0 \quad (9)$$

has a solution f^*, and

$$\operatorname*{argmin}_j f_j^*(x) = \operatorname*{argmin}_j p(y = j|x) \quad (10)$$

Proof: Firstly, we proof the existence of a minimizer f^*: Since p_i and $V(f_i)$ are both non-negative, the objectivefunction is bounded below by 0. Because $V(\cdot)$ is continuous,the objective function is also continuous. Therefore we cansafely say that the existence of a minimizerf^* for theoptimization problem is guaranteed.

Secondly, we show that (10) holds. Without loss of generality, we assume $p_1 > p_2 \geq \cdots \geq p_k > 0$, where$p_i = P(y_i = i|x_i)$. If we have a minimizer $\hat{f}$ with $\hat{f}_m > \hat{f}_1\ (m \neq 1)$, then $f^* = (\hat{f}_m, \hat{f}_2, \ldots, \hat{f}_{m-1}, \hat{f}_1, \hat{f}_{m+1}, \ldots, \hat{f}^k)$ isa better solution of (9). This contradicts the assumption that$\hat{f}$is a minimizer of (9). Therefore, $f_1^* = \max_{i=1,\ldots,k}\{f_i\}$.

Remark 2: According to the nonconvexity of smoothed 0-1loss function, the solution to (8) is not required to be unique. Actually, the uniqueness is not a necessary condition for Fisher consistency. So we safely relax Def 5 [15] by not requiring an unique minimizer. In the binary case, the minimizer is also not required to be unique [14].

Remark 3: The constraint $\Sigma_{i=1}^{k} f_j(x) = 0$ guarantees that if a point z is in the symmetric set R (defined by [16]) then sois any point obtained by interchanging any two coordinates of z.

Next we demonstrate the geometric features of Fisher consistency for different loss functions in multiclass cases.

Similar to the binary case, we fix x and denote $P(Y = y_i|X = x)$ by p_i, and we omit the argument in $f(x)$. Theconditional expected value can be written as

$$A(f) = \sum_{i=1}^{k} p_i V(f_i) \quad (11)$$

If we define the set R $\subseteq$ R^k as

$$R = \{(V(f_1), V(f_2), \ldots, V(f_k)) : f = (f_1, \ldots, f_k) \in R^k\}, \quad (12)$$

then the minimization of $A(f)$ can be written as

$$\min_{z \in R} \langle p, z \rangle \quad (13)$$

where $p = (p_1, p_2, \ldots, p_k)$.

The set R is shown in Figure 1(a) for the squared lossfunction $V(t) = (1 - t)^2$ by a blue curved surface. Givenp, geometrically, the solution to (13) is obtained by takinga hyper plane (denoted in Fig. 1(a) by green hyper plane) whose equation is $\langle p, z \rangle = c$ and then sliding it until it touches$R$ and with the minimum$c$.

It is intuitively clear from Figure 1(a) that if$p_1 > p_2 > p_3$, then the angle between the hyper plane and the axes $V(f_1)$ isthe biggest, and the value of $V(f_1)$ is the smallest compared with $V(f_2)$ and $V(f_3)$. Because $V(\cdot)$ is a decreasing

functionin $[-\infty, 1]$, we can safely say that

$$\underset{j}{\operatorname{argmin}} f_j(x) = \underset{j}{\operatorname{argmin}} p(y = j|x) \tag{14}$$

Now it is clear that the square loss function is Fisher consistent in this 3 dimensional case.

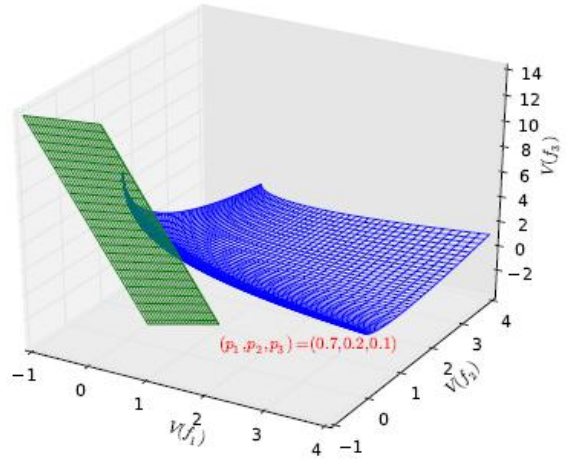

Fig. 1(a). The demonstration of geometric features of Fisher consistency for Square Loss Function.

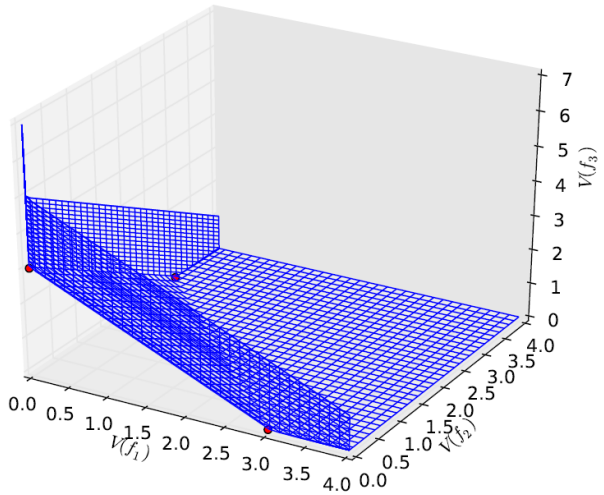

Fig. 1(b). The demonstration of geometric features of Fisher consistency for Hinge Loss Function.

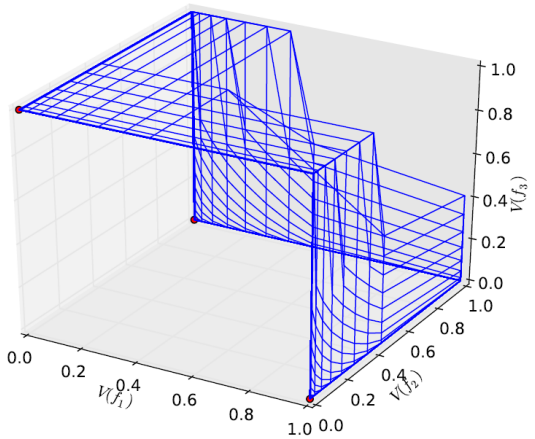

Fig. 1(c).The demonstration of geometric features of Fisher consistency for Smoothed 0-1 Loss Function.

In Fig. 1(b), the set R of hinge loss function is shown bya blue figure. $\forall p = (p_1, p_2, p_3)$, we can have a set of hyper planes. However, the hyper planes that minimize Equation (13) have to pass the three red points. Therefore, the solution of this problem is also unique, and Equation (14) is satisfied. The Fisher consistent condition is satisfied in this case. In Fig. 1(c), the set R of the smoothed 0-1 loss function is denoted byblue figures. The same as the square loss and hinge loss, thesolution to equation 14 is also unique. Therefore, the geometryexample shows that the smoothed 0-1 loss function satisfies Theorem 8.

IV. Single Machine Approach for Multiclass Data Classification

Based on the Tikhonov regularization model and the multiclass hinge loss function (2), we have the following multiclass classification model:

$$\min_{W,b} \frac{1}{l}\sum_{i=1}^{l}\sum_{j\neq y_i}\left(1-\left(f_{iy_i}-f_{ij}\right)\right)_+ + \lambda\sum_{j=1}^{k}\left||w_j|\right|^2 \tag{15}$$

By replacing the multiclass hinge loss $\left(f_{iy_i}-f_{ij}\right)_+$ byξ_i and making some simple transformation, we can have thefollowing model proposed by [17].

Considering smoothed 0-1 loss function, we propose the following single machine model for multiclass classification problems.

$$\min_{W,b} \phi(W,b) = \frac{1}{l}\sum_{i=1}^{l}\sum_{j\neq y_i} V(t_{ij}) + \lambda\sum_{j=1}^{k}\left||w_j|\right|^2 \tag{16}$$

where

$$V(t_j) = \begin{cases} 0, & t_j > 1 \\ \frac{1}{4}t_j^3 - \frac{3}{4}t\_j + \frac{1}{2}, & -1 \le t_j \le 1 \\ 1, & t_j < -1 \end{cases} \tag{17}$$

$$t_{ij} = f_{iy_i} - f_{ij} = \langle x_i, w_{y_i}\rangle + b_{y_i} - \langle x_i, w_j\rangle - b_j (j \neq y_i)$$

The derivative of objective function in (16) can be expressed as:

$$\frac{\partial\phi}{\partial(w_{ij})} = \begin{cases} \frac{1}{l}\sum_{i=1}^{l}\sum_{j\neq y_i} x_{ij}\left(0.75 - 0.75t_j^2\right) + 2\lambda w_{ij} \\ if\ j \neq y_i\ and -1 \le t_j \le 1 \\ 2\lambda w_{ij} \\ if\ j = y_i\ or\ t_j < -1\ or\ t_j > 1 \end{cases} \tag{18}$$

$$\frac{\partial\phi}{\partial(b_j)} = \begin{cases} \frac{1}{l}\sum_{i=1}^{l}\sum_{j\neq y_i}\left(0.75 - 0.75t_j^2\right) \\ if\ j \neq y_i\ and -1 \le t_j \le 1 \\ 0 \\ if\ j = y_i\ or\ t_j < -1\ or\ t_j > 1 \end{cases} \tag{19}$$

According to the above single machine model for multiclass classification problems, we propose the corresponding multiclass classification algorithm (see Algorithm 1).

It has been demonstrated that Quasi Secant Method (QSM) [18] outperforms some traditional local optimization methods for binary classification problem [13]. In this algorithm, QSMis adopted as the optimization solver. To apply the QSMin the multiclass case, we make a simple transformation by converting matrix W and vector b into a long vector.

Algorithm 1: The smoothed 0-1 loss function based multi-class training procedure

Input: $S = \big((x_1, y_1), \ldots, (x_l, y_l)\big), \Lambda, W, b$
while do
stoping criteria not satisfied
end
foreach$i = 1, \ldots, l$do
$f(x_i) \leftarrow x_i \cdot W + b$;
foreach$j = 1, \ldots, k$ and $j \neq y_i$do
$t_j = f_{y_i}(x_i) - f_j(x_i)$;
if $t_j > 1$ then
$V = 0$;
else if $1 \geq t_j \geq -1$ then
$V = 0.25 * t_{3i} - 0.75 * t_i + 0.5$;
else
$V = 1$;
end
$V_{total} = V_{total} + V$
end
end
foreach$i = 1, \ldots, k$do
$R = ||w_i||_2$;
$R_{total} = R_{total} + R$;
end
$Objf = V\ total/l + \lambda * R_{total}$;
Call optimization solver to minimize Objf ;
return W, b.

V. Numerical Experiments

In this Section, we examine the effectiveness of the proposed single machine classification algorithm for multiclass learning. Hinge loss function and the Smoothed 0-1 Loss function based algorithms are compared. We present experimental results of nine multiclass classification problems from the Statlog collection [19] and the UCI Repository of machine learning databases [20] respectively. From the UCI Repository we choose the following data sets: iris, wine, glass, vowel, ecoli, letter, and pendigit. From Statlog collection we choose the following data sets: vehicle, and segment. These multiclass classification problems had already been tested in [17], [21]-[23] respectively. The data set statistics are given in Table I. Some data sets in this table do not have a training and test split, in this case we use 10-fold cross validation approach to evaluate the performance of different classification algorithms.

TABLE I: Multiclass Data Sets Used in Numerical Experiments

Problem	#training data	#test data	#classes	#attributes
ecoli	327		5	7
iris	150	2100	3	4
wine	178		3	13
glass	214		6	13
vowel	990		11	10
vehicle	846		4	18
segment	210		7	19
pendigit	7494	3498	10	16
letter	15000	5000	26	16

In order to investigate the robustness of the proposed classification algorithm, random class and attribute noise are injected into these data sets respectively. These noisy data sets are generated by randomly corrupting 20% of training data, and we keep the test data intact.

For class noise, we corrupt 20% training data by updating the class label to a random class from all possible classes.

For attribute noise, instead of corrupting each attribute by a random value that is between the maximal and minimal [24], we corrupt the attribute by either the maximal or minimal of the corresponding attribute because this can help us insert outliers to the data set. In order to investigate the performance of robust classification algorithm, four different noise levels are considered for attribute noise. That is, 1 and 2 attribute values altered per data point; 50% and 100% attribute values altered per data point.

With this scheme, the actual percentage of noise is always lower than the theoretical noise level, as sometimes the random assignment would pick the original label.

The above mechanism implies that we only deal with completely random class or attribute noise, which means the probability that a label or an attribute has noise is unrelated to any other label or attribute. If noise among labels or attributes is introduced with correlations, the situation becomes more complicated, and this is beyond the coverage of this research.

The classification results of hinge loss Function and smoothed 0-1 loss function based algorithms on nine original data sets are presented in Table II. It can be seen that even though the training accuracy of smoothed 0-1 loss function based algorithm is much better, there is no clear improvement on the test accuracy. It is because these data sets are well behaved, in other words, these data sets do not have too much noise or outliers. To demonstrate the robustness of the proposed algorithm, next we investigate noisy data sets.

Table III lists the average CPU time over 100 runs. Even though the CPU time of smoothed 0-1 loss function based algorithm takes longer than hinge loss function based algorithm, it is still acceptable.

TABLE II: Comparison of Hinge Loss Function and the Smoothed 0-1 Loss Function on Multiclass Classification on Original Data Sets. Accuracies Are Presented for Training and Test Sets

	Hinge Loss		Smoothed0-1Loss		Improvement	
	training	test	training	test	training	test
ecoli	**86.97%**	**85.42**%	86.93%	85.06%	-0.04%	-0.36%
iris	96.47%	89.38%	**96.69%**	**91.25%**	0.22%	1.87%
wine	99.07%	90.90%	**98.02%**	**91.45%**	-1.05%	0.55%
glass	54.86%	**53.07%**	**55.94%**	52.15%	1.08%	-0.92%
vowel	48.44%	46.60%	**52.05%**	**47.80%**	3.61%	1.20%
vehicle	83.79%	**79.19%**	**85.14%**	78.48%	1.35%	-0.71%
segme	**99.05%**	**92.29%**	**99.05%**	91.91%	0.00%	-0.38%
pendigi	95.94%	**90.77%**	**97.40%**	89.54%	1.46%	-1.23%
letter	70.19%	70.17%	**73.05%**	**72.57%**	2.86%	2.40%

TABLE III: Comparison of Average CPU Time over 100 Runs

	Hinge Loss(s)	Smoothed0-1Loss(s)
ecoli	0.3985	0.8518
iris	0.0781	0.0906
wine	0.5844	0.8797
glass	0.0606	0.1153
vowel	8.2222	15.2194
vehicle	108.7729	114.1934
segment	8.1314	45.2806
pendigit	31.0965	50.2832
letter	171.5363	288.3232

Table IV gives the classification results on data sets with class noise. Compare to hinge loss function based classification algorithm, results of smoothed 0-1 loss function based classification algorithm in Table IV show that the improvements in classification performance for the smoothed 0-1 loss function based classification algorithm are significant for all training data sets and 8 test data sets

respectively. Even though the test accuracy on vowel data set is worse, the difference is not big (only 0.4%). It is also noticed that due to training sets are corrupted by label noise while test sets are intact, all training accuracy are less than the corresponding test accuracy.

TABLE IV: COMPARISON OF HINGE LOSS FUNCTION AND THE SMOOTHED 0-1 LOSS FUNCTION ON MULTICLASS CLASSIFICATION ON DATA SETS WITH 20% TRAINING LABEL NOISE. ACCURACIES ARE PRESENTED FOR TRAINING AND TEST SETS

	Hinge Loss		Smoothed0-1Loss		Improvement	
	training	test	training	test	training	test
ecoli	62.59%	70.73%	**65.84%**	**73.51%**	3.25%	2.78
iris	81.25%	88.13%	**84.63%**	**90.63%**	3.38%	2.50
wine	84.99%	86.41%	**85.54%**	**90.13%**	0.55%	3.72
glass	51.44%	55.40%	**53.77%**	**56.60%**	2.33%	1.20
vowel	36.00%	**40.00%**	**36.20%**	39.60%	0.20%	-0.40
vehicle	68.49%	74.19%	**72.56%**	**76.01**%	4.07%	1.82
segmen	69.19%	73.44%	**76.78%**	**85.25**%	7.59%	11.81
pendigi	74.54%	84.71%	**78.55%**	**87.48**%	4.01%	2.77
letter	52.80%	64.63%	**57.82%**	**71.35**%	5.02%	6.72

Table V and Table VI summarize the classification accuracy of the two algorithms on noisy data sets with 1 and 2 attributes are altered by maximal or minimal values of the corresponding attributes.

Compared to hinge loss function based algorithm, we see that smoothed 0-1 loss function based algorithm is effective in increasing classification accuracy, even though the improvement is not as prevalent as the results shown in Table IV with label noise. The reason why the attribute noise experimental results in Table V and Table VI are less significant than the label noise experiments shown in Table IV is that there are only one or two attributes are altered per data point, these attributes may not be predictable attributes and accordingly degrade the noise level. To see a more significant attribute noise level, we investigate Table VII and Table VIII where50% and 100% attributes are corrupted by noise.

TABLE V: COMPARISON OF HINGE LOSS FUNCTION AND THE SMOOTHED 0-1 LOSS FUNCTION ON MULTICLASS CLASSIFICATION ON DATA SETS WITH 20% TRAINING ATTRIBUTE NOISE (ONE ATTRIBUTE VALUE ALTERED PER DATA POINT). ACCURACIES ARE PRESENTED FOR TRAINING AND TEST SETS

	Hinge Loss		Smoothed0-1Loss		Improvement	
	training	test	training	test	training	test
ecoli	80.83%	**83.65**%	**81.61%**	82.55%	0.78%	-1.10%
iris	90.51%	**88.75**%	**90.96%**	**88.75%**	0.45%	0.00%
wine	**97.33%**	90.51%	96.64%	**91.07%**	-0.69%	0.56%
glass	51.24%	48.62%	**54.24%**	**50.47%**	3.00%	1.85%
vowel	42.80%	43.90%	**44.22%**	**45.00%**	1.42%	1.10%
vehicle	70.00%	**73.93**%	**79.18%**	73.72%	9.18%	-0.21%
segment	**99.05%**	**92.19**%	**99.05%**	90.77%	0.00%	-1.42%
pendigit	90.55%	88.45%	**93.42%**	**89.00**%	2.87%	0.55%
letter	62.91%	68.87%	**64.52%**	**70.79**%	1.61%	1.92%

TABLE VI: COMPARISON OF HINGE LOSS FUNCTION AND THE SMOOTHED 0-1 LOSS FUNCTION ON MULTICLASS CLASSIFICATION ON DATA SETS WITH 20% TRAINING ATTRIBUTE NOISE (TWO ATTRIBUTE VALUE ALTERED PER DATA POINT). ACCURACIES ARE PRESENTED FOR TRAINING AND TEST SETS.

	Hinge Loss		Smoothed0-1Loss		Improvement	
	training	test	training	test	training	test
ecoli	76.60%	**80.67%**	**77.15%**	78.91%	0.55%	-1.76%
iris	88.38%	**88.75%**	**89.63%**	**88.75%**	1.25%	0.00%
wine	95.60%	**90.17**%	**97.03%**	89.40%	1.43%	-0.77%
glass	49.33%	46.38%	**53.41%**	**46.84%**	4.08%	0.46%
vowel	40.92%	42.60%	**42.79%**	**43.40**%	1.87%	0.80%
vehicle	68.49%	74.19%	**72.56%**	**76.01**%	4.07%	1.82%
segment	96.68%	**89.86**%	**98.10%**	89.77%	1.42%	-0.09%
pendigit	87.48%	87.40%	**90.94%**	**88.91%**	3.46%	1.51%
letter	60.46%	69.15%	**62.64%**	**71.63**%	2.18%	2.48%

TABLE VII: COMPARISON OF HINGE LOSS FUNCTION AND THE SMOOTHED 0-1 LOSS FUNCTION ON MULTICLASS CLASSIFICATION ON DATA SETS WITH 20% TRAINING ATTRIBUTE NOISE (50% ATTRIBUTE VALUES ALTERED PER DATA POINT). ACCURACIES ARE PRESENTED FOR TRAINING AND TEST SETS

	Hinge Loss		Smoothed0-1Loss		Improvement	
	training	test	training	test	trainin	test
ecoli	**76.23%**	80.37%	75.76%	**80.67**%	-0.47	0.3%
iris	87.57%	**88.75**%	**89.34%**	**88.75%**	1.77%	0%
wine	93.60%	90.90%	**94.16%**	**91.62%**	0.56%	0.72%
glass	48.29%	45.58%	**50.31%**	**47.40%**	2.02%	1.82%
vowel	36.61%	39.80%	**40.57%**	**41.30%**	3.96%	1.5%
vehicle	66.11%	68.72%	**72.90%**	**74.07%**	6.79%	5.35%
segment	97.16%	86.48%	**98.58%**	**87.77%**	1.42%	1.29%
pendigit	78.81%	84.48%	**82.62%**	**87.54**%	3.81%	3.06%
letter	54.06%	65.09%	**59.54%**	**70.97**%	5.48%	5.88%

Table VII and Table VIII show the results of experiments with higher attribute noise levels. First, except for data set "ecoli" and "iris", smoothed 0-1 loss function based algorithm out performs hinge loss function based algorithm on both training and test accuracy. Second, it is clear that with the increasing of the attribute noise level (from 50% to 100%), smoothed 0-1 loss function based algorithm performs better on test accuracy. The only exception cases are data set "iris" and "vehicle", all other seven data sets' test classification accuracy have been better increased. Moreover, for data set "vowel" and "letter", smoothed 0-1 loss function based algorithm is even able to increase the classification accuracy significantly by 6.7% and 6.82% respectively when the attribute noise level as high as 100% in Table VIII. These experiments also demonstrate that in this case label noise has a larger effect or impact on the performance of the classifiers than attribute noise.

TABLE VIII: COMPARISON OF HINGE LOSS FUNCTION AND THE SMOOTHED 0-1 LOSS FUNCTION ON MULTICLASS CLASSIFICATION ON DATA SETS WITH 20% TRAINING ATTRIBUTE NOISE (100% ATTRIBUTE VALUES ALTERED PER DATA POINT). ACCURACIES ARE PRESENTED FOR TRAINING AND TEST SETS

	Hinge Loss		Smoothed0-1Loss		Improvement	
	training	test	training	test	training	test
ecoli	72.51%	78.36%	**72.87%**	**79.27%**	0.36%	0.91%
iris	84.49%	**90.63%**	**87.50%**	90.00%	3.01%	-0.63%
wine	88.82%	85.68%	**90.50%**	**89.79%**	1.68%	4.11%
glass	46.08%	41.49%	**50.21%**	**44.22%**	4.13%	2.73%
vowel	34.30%	37.90%	**40.49%**	**44.60**%	6.19%	6.7%
vehicle	64.43%	68.68%	**72.21%**	**71.34**%	7.57%	2.66%
segment	93.37%	87.34%	**95.26%**	**89.24%**	1.89%	1.9%
pendigit	74.13%	82.60%	**79.59%**	**86.05%**	5.46%	3.45%
letter	52.30%	63.81%	**58.45%**	**70.63**%	6.15%	6.82%

VI. CONCLUSION

This paper extends smoothed 0-1 loss function from binary to multiclass classification problem. From the theoretical point of view, Fisher consistency of smoothed 0-1 loss function has been proved and demonstrated by geometrical examples. A multiclass classification algorithm is proposed based on a single machine model. These algorithms are used to demonstrate the robustness of smoothed 0-1 loss function based classification algorithm. For experiments, 9 multiclass datasets are used and different kinds of noise with different levels are introduced to these data sets. The experiments show that smoothed 0-1 loss function based multiclass classification algorithm is robust.

ACKNOWLEDGMENTS

The authors gratefully thank Adil Bagirov for his advice

and support, particularly for his excellent Quasisecant optimization method and corresponding Fortran source codes.

REFERENCES

[1] R. Rifkin, "Everything old is new again: A fresh look at," Ph.D. thesis, Massachusetts Institute of Technology, 2002.

[2] Y. Freund and R. Schapire, "A decision-theoretic generalization of online learning and an application to boosting," *Computational Learning Theory*, pp. 23–37, Springer, 1995.

[3] S. Agarwal, "Surrogate regret bounds for the area under the roc curve via strongly proper losses," *COLT*, 2013.

[4] J. Friedman, T. Hastie, and R. Tibshirani, "Special invited paper. Additive logistic regression: A statistical view of boosting," *Annals of Statistics*, vol. 28, no. 2, pp. 337–374, 2000.

[5] C. Domingo and O. Watanabe, "Mada Boost: A modification of Ada Boost," in *Proc. the Thirteenth Annual Conference on Computational Learning Theory*, Citeseer, 2000, pp. 180–189.

[6] L. Mason, P. Bartlett, and J. Baxter, "Improved generalization through explicit optimization of margins," *Machine Learning*, vol. 38, no. 3, pp. 243–255, 2000.

[7] X. Shen, G. Tseng, X. Zhang, and W. Wong, "On -learning," *Journal of the American Statistical Association*, vol. 98, no. 463, pp. 724–734, 2003.

[8] R. Collobert, F. Sinz, J. Weston, and L. Bottou, "Large scale transductive SVMs," *Journal of Machine Learning Research*, vol. 7, pp. 1687–1712, 2006.

[9] R. Collobert, F. Sinz, J. Weston, and L. Bottou, "Trading convexity for scalability," in *Proc. the 23rd International Conference on Machine Learning*, ACM, 2006, pp. 179–208.

[10] Y. Wu and Y. Liu, "Robust truncated hinge loss support vector machines," *Journal of the American Statistical Association*, vol. 102, no. 479, 2007.

[11] L. Mason, J. Baxter, P. Bartlett, and M. Frean, "Functional gradient techniques for combining hypotheses," *Advances in Neural* Information *Processing Systems*, pp. 221–246, 1999.

[12] N. Krause and Y. Singer, "Leveraging the margin more carefully," in *Proc. the 21st International Conference on Machine Learning*, ACM, 2004, pp. 63.

[13] L. Zhao, M. Mammadov, and J. Yearwood, "From convex to nonconvex: A loss function analysis for binary classification," in *Proc. 2010 IEEE International Conference on Data Mining Workshops*, IEEE, 2010, pp. 1281–1288.

[14] Y. Lin, "A note on margin-based loss functions in classification," *Statistics & Probability Letters*, vol. 68, no. 1, pp. 73–82, 2004.

[15] H. Zou, J. Zhu, and T. Hastie, "New multicategory boosting algorithms based on multicategory Fisher-consistent losses," *The Annals of Applied Statistics*, vol. 2, no. 4, pp. 1290–1306, 2008.

[16] A. Tewari and P. Bartlett, "On the consistency of multiclass classification methods," *The Journal of Machine Learning Research*, vol. 8, pp. 1007–1025, 2007.

[17] J. Weston and C. Watkins, "Multi-class support vector machines," *Pattern Recognition*, 1998.

[18] A. Bagirov and A. Ganjehlou, "A quasisecant method for minimizing nonsmooth functions," *Optimization Methods and Software*, vol. 25, no. 1, pp. 3–18, 2010.

[19] D. Michie, D. Spiegelhalter, C. Taylor, and J. Campbell, *Machine Learning, Neural and Statistical Classification*, Ellis Horwood, 1995.

[20] A. Asuncion and D. Newman, "UCI machine learning repository," University of California, Irvine, School of Information and Computer Sciences, 2010.

[21] V. Vapnik, *Statistical Learning Theory*, Wiley-Interscience, 1998.

[22] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *The Journal of MachineLearning Research*, vol. 2, pp. 265–292, 2002.

[23] C. Hsu and C. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[24] X. Zhu and X. Wu, "Class noise vs. attribute noise: A quantitative study," *Artificial Intelligence Review*, vol. 22, no. 3, pp. 177–210, 2004.

Lei Zhao is a research information coordinator in the Office of Research at University of the Sun-shine Coast. He completed his Ph.D at University of Ballarat in 2012.