Robust Boosting Algorithm Against Mislabling in Multiclass Problems

Takashi Takenouchi
ttakashi@is.naist.jp
Nara Institute of Science and Technology, Graduate School of Information Science, Ikoma, Nara 630-0192, Japan

Shinto Eguchi
eguchi@ism.ac.jp
Institute of Statistical Mathematics, Japan and Department of Statistical Science, Graduate University of Advanced Studies, Minami-azabu, Tokyo 106-8569, Japan

Noboru Murata
noboru.murata@eb.waseda.ac.jp
Faculty of Science and Engineering, Waseda University, Shinjuku, Tokyo 169-8555, Japan

Takafumi Kanamori
kanamori@is.nagoya-u.ac.jp
Department of Computer Science and Mathematical Informatics, Graduate School of Information Science, Nagoya University, Furocho, Chikusaku, Nagoya 464-8603, Japan

We discuss robustness against mislabeling in multiclass labels for classification problems and propose two algorithms of boosting, the normalized Eta-Boost.M and Eta-Boost.M, based on the Eta-divergence. Those two boosting algorithms are closely related to models of mislabeling in which the label is erroneously exchanged for others. For the two boosting algorithms, theoretical aspects supporting the robustness for mislabeling are explored. We apply the proposed two boosting methods for synthetic and real data sets to investigate the performance of these methods, focusing on robustness, and confirm the validity of the proposed methods.

1 Introduction

Classification and pattern recognition in statistical learning theory is now heading in a new direction. A number of learning algorithms of boosting have been proposed since AdaBoost was introduced in Freund and Schapire (1997), and their efficiency has been elucidated for classification problems from various points of view (Schapire & Singer, 1999; Eibl & Pfeiffer, 2005).
The idea is based on combining different learning machines to build a high-performance machine.

In this letter, we introduce a new probabilistic formalization for multiclass classification algorithms. In principle, probabilistic formalization covers various situations connecting feature vectors with labels in accordance with the degree of random variability. There are different strategies for dealing with different types of classification problems. For example, precision goals depend on data contexts in problems for classifying handwritten characters or printed ones in character recognition. We focus on the classification based on empirical examples with existing noise in labeling feature vectors. Regarding the binary classification problem, a contamination model of a conditional probability of the label for the feature vector is introduced in Takenouchi and Eguchi (2004). Therefore, the practical conditional probability is considered to be a mixture of the assumed probability distribution and the incorrect probability due to mislabeling.

In this letter, we propose Eta-Boost.M for multiclass classification as an extension of the binary \( \eta \)-Boost algorithm (Takenouchi & Eguchi, 2004). Eta-Boost.M consists of a single-parameter family of boosting algorithms with parameter \( \eta \), which subsumes logistic discriminant analysis with \( \eta = 0 \). This is based on Eta-divergence between densities, which need not have a total mass of 1. The Kullback-Leibler divergence yields a pair of boosting algorithms, AdaBoost and LogitBoost (Lebanon & Lafferty, 2001). According to this analogy, Eta-divergence has a pair of boosting algorithms, Eta-Boost.M, and normalized Eta-Boost.M, associated with two models of mislabeling. The key idea is to focus on a uniform distribution on the labels. The classification task becomes more difficult when the conditional probability distribution, given a feature vector, becomes closer to the uniform distribution. From this point of view, we propose a mislabeling model for the conditional probability such that the probability of a mislabel is relatively high when the supposed conditional probability distribution, given the feature vector, is faced with a difficult situation as described above. This is a probabilistic extension of the mislabeling model discussed in Takenouchi and Eguchi (2004). We discuss several versions of eta-loss functions associated with the mislabeling model.

2 Probabilistic Model for Mislabeling

2.1 Formulation. Let us take a probabilistic formulation for a classification problem. Let \( X \) be a feature vector in a space \( \mathcal{X} \) and \( Y \) be a label in a finite set \( \mathcal{Y} = \{1, \ldots, G\} \). Suppose that an underlying probability distribution of \((X, Y)\) is

\[
    r(x, y) = q(x)p(y|x),
\]

and we introduce a formal discussion in an idealized situation where the distribution \( r(x, y) \) is completely known. Thus, we have complete
knowledge of the conditional probability of $Y = y$ given $X = x$, say, $p(y|x)$. Then the Bayes rule provides the optimal rule of classification for a given feature vector $x$:

$$F_p^{\text{bayes}}(x) = \arg\max_{y \in \mathcal{Y}} p(y|x).$$

(2.2)

The sense of optimality of $F_p^{\text{bayes}}$ is justified by minimizing the error rate of a classification rule $F(x)$ in the class of all classification rules, and the error rate is defined by

$$\text{Err}(F, p, q) = \text{Prob}_{p, q}(F(X) \neq Y)$$

$$= 1 - \sum_{y \in \mathcal{Y}} \int_{X_{F, y}} p(y|x)q(x) \, dx,$$

where $X_{F, y} = \{x | x \in \mathcal{X}, F(x) = y\}$. As the optimality measure, the error rate can be extended to various criteria, including the credit score and the area under the ROC (receiver operating characteristic) curve. (See Eguchi & Copas, 2002, for a discussion of the relevant statistics.) Let $H(x, y)$ be a discriminant function associated with a classification rule $F_H(x)$ by

$$F_H(x) = \arg\max_{y \in \mathcal{Y}} H(x, y).$$

(2.3)

In a class of discriminant functions, $H$ is said to be Bayes optimal if $F_H = F_p^{\text{bayes}}$.

Let $f(x) \in \mathcal{Y}$ be a weak classification rule that needs a relatively low computational cost to be constructed. Note that $f(x)$ is a multiple value function, and it returns a subset of $\mathcal{Y}$. For a given classification rule $f(x)$, we define a weak hypothesis by

$$h_f(x, y) = I(y \in f(x)).$$

(2.4)

The weak hypothesis $h_f(x, y)$ is a kind of a discriminant function corresponding to the classification rule $f(x)$, and let $\mathcal{H}$ be the class of all weak hypotheses. The aim of a boosting algorithm is to construct a discriminant function by linearly combining $T$ weak hypotheses as

$$H_T(x, y) = \sum_{i=1}^{T} \alpha_i h_i(x, y), h_i \in \mathcal{H}.$$

2.2 Mislabeling Model. In many cases of classification problems, the class membership probability $p(y|x)$ is modeled by the conventional logistic
model \( p_{0,H}(y|x) \) constructed by the discriminant function \( H(x, y) \) as

\[
p_{0,H}(y|x) = \frac{\exp(H(x, y))}{\sum_{y' \in \mathcal{Y}} \exp(H(x, y'))},
\]

and the discriminant function \( H \) is estimated based on a given training data set \( \{(x_i, y_i), i = 1, \ldots, n\} \). In practice, the class labels in the data set are frequently determined by human judgments with varying degrees of uncertainty. Failure prediction and medical diagnostics are typical examples of such uncertainty. Thus, we often face a situation in which some of the labels of the given data set are wrongly observed and then have to consider a model in which some contamination may be added to the data generation or an observation mechanism in the labeling process. In a context of the binary classification problem, Copas (1988) considered a contamination model with a constant probability and discussed statistical properties. The model was extended to consider a probability of contamination depending on the input \( x \) with a boosting method algorithm (Takenouchi & Eguchi, 2004), and it was confirmed that an algorithm considering mislabeling worked well compared with other methods, such as the original AdaBoost, MadaBoost (Domingo & Watanabe, 2000), and AdaBoost_{reg} (Rätsch, Onoda, & Müller, 2001).

In this letter, we extend the concept of the contamination model to the multiclass classification problem. We consider a probabilistic approach to mislabeling problems as follows. Let us define

\[
\mathcal{D} = \left\{ \delta(x) \left| x \in \mathcal{X}, 0 \leq \delta(x) < \frac{1}{G} \right. \right\}
\]

and consider the following model:

\[
p_{\delta}(y|x) = (1 - (G - 1)\delta(x))p(y|x) + \sum_{k \neq y} \delta(x) p(k|x).
\]

This is a model of the posterior distribution of the class label \( y \) given the feature vector \( x \), such that a class labeling with the original posterior distribution \( p(y|x) \) is changed by mislabeling from \( y \) to \( G - 1 \) other labels with equal probability \( \delta(x) \). We call \( p_{\delta}(y|x) \) a mislabeled model with mislabeled probability \( \delta(x) \). It will be shown that the Bayes rule based on \( p_{\delta}(y|x) \) is invariant with that based on the original posterior \( p(y|x) \), while a lower bound for error rate increases for any \( \delta(x) \).

In classification problems, one of the most difficult situations is suggested by a case in which the posterior distribution does not depend on any feature vector \( x \), that is, \( p(y|x) = p_{U}(y) = 1/G \), where \( p_{U} \) is a uniform distribution on \( \mathcal{Y} \). If we choose a probability of mislabeled setting by \( \delta(x) = 1/G \), then we
observe that \( p_\delta(y|x) = p_U(y) \) and the model \( p_\delta(y|x) \) has no prediction ability. We will propose a more sensible form of \( \delta(x) \) in a subsequent discussion.

We first remark from equation 2.6 that

\[
p_\delta(y|x) = (1 - G\delta(x)) p(y|x) + \delta(x),
\]

which implies that the Bayes rule based on the original posterior \( p(y|x) \) is the same as that based on the current posterior \( p_\delta(y|x) \) for any \( \delta(x) \in \mathcal{D} \).

Remark 1. For any \( \delta \in \mathcal{D} \), we observe \( F^\text{bayes}_p(x) = F^\text{bayes}_{p_\delta}(x) \).

Proof. For any \( y, y' \in \mathcal{Y} \) and \( x \in \mathcal{X} \), we observe

\[
p_\delta(y|x) - p_\delta(y'|x) = (1 - G\delta(x))(p(y|x) - p(y'|x)).
\]

Because \( \delta(x) \) is in \( \mathcal{D} \), \( 1 - G\delta(x) \) is positive, and then order relations of posteriors of \( y \) are invariant.

Second, we investigate lower bounds of error rates for both cases of \( p(y|x) \) and \( p_\delta(y|x) \). Let us consider a lower bound \( \text{errbd}(p, q) \) of the error rate under \( p(y|x) q(x) \),

\[
\text{errbd}(p, q) = \text{Err}(F^\text{bayes}_p, p, q).
\]

Then we obtain the following theorem:

Theorem 1. For any \( \delta \in \mathcal{D} \), we observe

\[
\text{errbd}(p_\delta, q) \geq \text{errbd}(p, q).
\]

Proof. From remark 1, a region \( \mathcal{X}_{F^\text{bayes}_p, y} \) coincides with \( \mathcal{X}_{F^\text{bayes}_{p_\delta}, y'} \) and then we observe

\[
\text{errbd}(p_\delta, q) - \text{errbd}(p, q) = \sum_{y \in \mathcal{Y}} \int_{\mathcal{X}^\text{bayes}_{F^\text{bayes}_p, y}} \{ p(y|x) - p_\delta(y|x) \} q(x) \, dx.
\]

Because of equation 2.7, we obtain

\[
\text{errbd}(p_\delta, q) - \text{errbd}(p, q) = \sum_{y \in \mathcal{Y}} \int_{\mathcal{X}^\text{bayes}_{F^\text{bayes}_p, y}} G\{ p(y|x) - p_U(y) \} \delta(x) q(x) \, dx.
\]

(2.10)
Here we observe for any class label \( y \in \mathcal{Y} \) that

\[
x \in \mathcal{X}_{F^{\text{bayes}}_p \cdot y} \implies p(y|x) = \max_{k \in \mathcal{Y}} p(k|x) \geq p_U(y).
\]

Hence, any integrand in equation 2.10 is nonnegative, which completes the proof.

Finally, we consider which form of the mislabeled probability \( \delta(x) \) is appropriate. In practice situation, it is conceivable that a failure of labeling in the observation process is likely to occur in a region where the classification based on the original posterior \( p(y|x) \) is difficult. Then it is necessary to model \( \delta(x) \) in accordance with the difficulty of the classification task. For this purpose, it is natural to consider the following situation: if the feature vector \( x \) is in a region of \( \mathcal{X} \) in which a maximum of \( p(y|x) \) on \( y \in \mathcal{Y} \) is almost 1, then \( \delta(x) \) should be smaller, while if \( x \) is in a region of \( \mathcal{X} \) in which the maximum is almost \( 1/G \), then \( \delta(x) \) should be larger. The easiest case is suggested when \( p(y|x) \) becomes \( I(y = F^{\text{bayes}}_p(x)) \), and the most difficult case is suggested when \( p(y|x) \) becomes \( p_U(y) \). To quantify a degree of the classification easiness, we use the Kullback-Leibler divergence between \( p_U(y) \) and \( p(y|x) \),

\[
d_Y(p_U, p(\cdot|x)) = \sum_{y \in \mathcal{Y}} p_U(y) \log \frac{p_U(y)}{p(y|x)} = \log \frac{1}{G} - \frac{1}{G} \sum_{y \in \mathcal{Y}} \log p(y|x),
\]

which measures the discrepancy between distributions and satisfies that

\[
p(\cdot|x) \to p_U \implies d_Y(p_U, p(\cdot|x)) \to 0,
\]

and

\[
p(\cdot|x) \to I(y = F^{\text{bayes}}_p(x)) \implies d_Y(p_U, p(\cdot|x)) \to \infty.
\]

In section 4, we propose a learning algorithm associated with a mislabeling probability

\[
\delta(x) = v(d_Y(p_U, p(\cdot|x))),
\]

where \( v \) is a monotonically decreasing fixed function such that \( v(0) \leq 1/G \) and \( \lim_{t \to \infty} v(t) = 0 \).
While the discussion here is consistently based on model 2.6, the model is rather restrictive. For example, we can extend to the model to consider a situation in which a mislabeling probability depends on the class label \( y \) as follows:

\[
p_{\Delta}(y|x) = Z(x)^{-1}\left\{ \left( 1 - \sum_{k \neq y} \delta_k(x) \right) p(y|x) + \sum_{k \neq y} \delta_k(x) p(k|x) \right\}
\]

where \( \Delta = (\delta_1, \ldots, \delta_G) \), \( \delta_k(x) \in \mathcal{D} \) is a probability of mislabeling in which the label is erroneously observed as \( k \) and

\[
Z(x) = \left( 1 - \sum_{k \in \mathcal{Y}} \delta_k(x) \right) + G \sum_{k \in \mathcal{Y}} \delta_k(x) p(k|x).
\]

This model would be much more complex and flexible than model 2.6; however, in this letter, we focus on model 2.6.

3 Bregman \( U \)-Loss Function and Constant Volume Bias Condition

Throughout this letter, we assume that the probability density function of the feature vector \( x \) is fixed to \( q(x) \) and \( \tilde{q}(x) \) is its empirical version. Let

\[
\mathcal{P} = \left\{ m(y|x) \left| x \in \mathcal{X}, m(y|x) > 0, \sum_{y \in \mathcal{Y}} m(y|x) = 1 \right. \right\}
\]

be the space of all the conditional probability distributions of \( Y \) given \( X = x \), and let

\[
\mathcal{M} = \left\{ m(y|x) \left| x \in \mathcal{X}, m(y|x) > 0, \sum_{y \in \mathcal{Y}} m(y|x) < \infty \right. \right\}
\]

be the extended space of \( \mathcal{P} \) with positive finite mass. Thus, \( m(y|x) \) in \( \mathcal{M} \) does not always satisfy \( \sum_{y \in \mathcal{Y}} m(y|x) = 1 \).

In this formulation, we take a convex and monotonically increasing function \( U \) on the real line. For \( m, \mu \in \mathcal{M} \), the Bregman \( U \)-divergence \( D_U \) defined
over $M \times M$ is given by

$$D_U(m, \mu) = \int_X q(x) \sum_{y \in Y} \Phi(m, \mu) \, dx.$$  

The function $\Phi(a, b)$ is defined as

$$\Phi(a, b) = U(u^{-1}(b)) - U(u^{-1}(a)) - a\{u^{-1}(b) - u^{-1}(a)\},$$

where $u(z) = U'(z)$ and $u^{-1}$ is the inverse function of $u$. From the convexity of $U$, $\Phi(a, b) \geq 0$ for any $a > 0, b > 0$. We observe that $D_U(m, \mu) \geq 0$ by definition. Also, if we set $U(z) = \exp(z)$, $U$-divergence is reduced to the extended Kullback-Leibler divergence.

We consider a pseudoconditional probability $\mu(y|x)$ in $M$, which is connected with a discriminant function $H(x, y)$ as

$$\mu(y|x) = u(H(x, y) - b_H(x)),$$  

(3.3)

where $b_H(x)$ is a bias function depending on the $H(x, y)$ but not on the label $y$. In Murata, Takenouchi, Kanamori, and Eguchi (2004), we proposed bias functions satisfying the normalized condition and the moment matching condition, and investigated statistical properties. The function $u(z)$, the derivative of $U(z)$, is the positive monotone increasing function from the convexity and monotonicity of $U(z)$, and the classification rule associated with $H(x, y)$ is not changed by $u(z)$ or any choice of $b_H(x)$; that means

$$\arg\max_{y \in Y} H(x, y) = \arg\max_{y \in Y} u(H(x, y) - b(x)).$$

By substituting $m(y|x) = p(y|x)$ and $\mu(y|x) = u(H(x, y) - b_H(x))$ into $D_U(m, \mu)$, respectively, and omitting terms that do not affect an optimization of $H(x, y)$, we obtain the $U$-loss function for the discriminant function $H(x, y)$ as

$$L_U(H) = \int_X q(x) \sum_{y \in Y} \{U(H(x, y) - b_H(x)) - p(y|x)(H(x, y)$$

$$- b_H(x))\} \, dx.$$  

(3.4)

(See Murata et al., 2004, for detailed discussions.) In fact, this $U$-loss function is equipped with a reasonable property in which the loss has a minimum when $H(x, y)$ generates the Bayes rule. Thus, the function $U$ generates the $U$-divergence and the $U$-loss function $L_U$. Then a variety of the $U$-loss functions corresponds to that of the function $U$. An empirical version of the
loss, equation 3.4, and a general algorithm derived from it, are shown in appendix A.

In the next section, we propose a new bias function, a constant volume condition, which produces a statistical model associated with the function \( U \).

### 3.1 Condition of the Bias Function and Associative Model.

Let us discuss the class of optimal discriminant functions that is obtained by minimizing the given \( U \)-loss function, equation 3.4, associated with three types of conditions for the bias function \( b_H(x) \). For the first condition, we obtain the following theorem:

**Theorem 2 (Constant volume condition).** Let \( H^* \) be a minimizer of \( U \)-loss function \( L_U(H) \) under the condition

\[
\sum_{y \in Y} U(H(x, y) - b_H(x)) = C, \tag{3.5}
\]

where \( C \) is a constant. Then \( H^* \) satisfies

\[
p(y|x) = \frac{u(H^*(x, y) - b_{H^*}(x))}{\sum_{y' \in Y} u(H^*(x, y') - b_{H^*}(x))}, \tag{3.6}
\]

and hence \( H^* \) is Bayes optimal.

**Proof.** We seek a minimizer of equation 3.4 under constraint 3.5 and find that it is sufficient to minimize equation 3.4 conditional on \( x \) (cf. Friedman, Hastie, & Tibshirani, 2000). By setting a variational derivative of equation 3.4 conditional on \( x \) to 0, we observe

\[
0 = \frac{\partial}{\partial H(x, y)} \sum_{y' \in Y} [U(H(x, y') - b_H(x)) - p(y'|x)(H(x, y') - b_H(x))] \bigg|_{H=H^*} = -p(y|x) + u(H^*(x, y) - b_{H^*}(x))
\]

\[
+ \left( 1 - \sum_{y' \in Y} u(H^*(x, y') - b_{H^*}(x)) \right) \frac{\partial b_{H^*}(x)}{\partial H(x, y)} \bigg|_{H=H^*}. \tag{3.7}
\]

Also, we observe the following relation from a variational derivative of condition 3.5:

\[
\forall y \in Y, \quad \frac{\partial b_H(x)}{\partial H(x, y)} = \frac{u(H(x, y) - b_H(x))}{\sum_{y' \in Y} u(H(x, y') - b_H(x))}. \tag{3.8}
\]
By substituting equation 3.8 into equation 3.7, we conclude equation 3.6. The denominator of the right-hand side of equation 3.6 does not depend on \( y \), and the function \( u \) is the monotonically increasing one. Thus, \( H^* \) is Bayes optimal.

We describe condition 3.5 as the constant volume condition. Note that we shall denote the object under condition 3.5 by one with a superscript \( v \) as a \( U \)-loss function \( L_U^v \) or a bias function \( b^v_H \). The constant value \( C \) is conventionally fixed to \( GU(0) \), where we impose \( b_H(x) = 0 \) as \( H(x, 1) = \cdots = H(x, G) = 0 \). Under condition 3.5, the \( U \)-loss function is reduced to

\[
L_U^v(H) = C - \int_X q(x) \sum_{y \in Y} p(y|x)(H(x, y) - b^v_H(x)) \, dx. \tag{3.9}
\]

We will omit the first term \( C \) since the optimization of the discriminant function \( H \) does not depend on the term. An algorithm is derived from an empirical version of equation 3.9 as in the case of appendix A, and we note an error rate property of the algorithm under the constant volume condition in appendix B.

For the second and third conditions, detailed properties have been previously investigated in Murata et al. (2004), and we observe the following corollaries:

**Corollary 1 (Normalized condition).** Let \( H^* \) be a minimizer of \( U \)-loss function \( L_U(H) \) under the following conditions:

\[
\sum_{y \in Y} u(H(x, y) - b_H(x)) = 1. \tag{3.10}
\]

Then \( H^* \) satisfies equation 3.6 and is Bayes optimal.

**Proof.** From equation 3.7 and condition 3.10, we immediately obtain the relation

\[
p(y|x) = u(H^*(x, y) - b_{H^*}(x)) = \frac{u(H^*(x, y) - b_{H^*}(x))}{\sum_{y' \in Y} u(H^*(x, y') - b_{H^*}(x))}. \tag{3.11}
\]

Furthermore, this relationship shows that \( H^* \) is Bayes optimal.
Corollary 2 (Moment matching condition). Let $H^*$ be a minimizer of $U$-loss function $L_U(H)$ under the conditions

$$
\sum_{y \in Y} p(y|x)(H(x, y) - b_H(x)) = 0.
$$

(3.12)

Then $H^*$ satisfies equation 3.6 and is Bayes optimal.

Proof. By differentiating condition 3.12, we observe

$$
0 = \frac{\partial}{\partial H(x, y)} \sum_{y' \in Y} p(y'|x)(H(x, y') - b_H(x)) \bigg|_{H=H^*} = p(y|x) - \frac{\partial b_H(x)}{\partial H(x, y)} \bigg|_{H=H^*}.
$$

(3.13)

By substituting equation 3.13 into 3.7, we conclude that $H^*$ satisfies equation 3.6 implying the Bayes optimality of $H^*$.

If we set $U(z) = \exp(z)$, the right-hand side of equation 3.6 with those three types of bias functions reduces to be the logistic model, which is used in the context of conventional statistical classification.

We describe condition 3.10 as a normalized condition and condition 3.12 as a moment matching condition. Details of the properties of those two bias conditions in the context of the boosting algorithm were discussed in Murata et al. (2004). The normalized condition constrains the pseudoconditional model, equation 3.3, in $\mathcal{M}$ to be the probabilistic model in $\mathcal{P}$, and derives the conventional logistic discriminant analysis with the function $U(z) = \exp(z)$. It is pointed out that the original AdaBoost is derived from the minimization of the extended Kullback-Leibler divergence between the empirical distribution and the extended exponential model in $\mathcal{M}$ under the moment matching condition (see Lebanon & Lafferty, 2001). This condition is also used in statistical inference. In the following context, we denote an object under the normalized condition 3.10 with a superscript $n$ and an object under the moment matching condition 3.12 with a superscript $m$, as $L^n_U$, $b^n_H$, $L^m_U$, and $b^m_H$, respectively. The $U$-loss function $L^m_U$ is written as

$$
L^m_U(H) = \int_{\mathcal{X}} q(x) \sum_{y \in Y} U(H(x, y) - b^m_H(x)) \, dx.
$$

3.2 Consistency for the Binary Case. For the binary classification problem, we can assume that the label $y$ is in $\{1, -1\}$ for convenience of calculation, and then the discriminant function $H(x, y)$ is represented by
$H(x, y) = F(x)y$, where a label is predicted by a sign of the function $F(x)$, $\text{sgn}(F(x))$. Then the $U$-loss function, equation 3.4, is rewritten as

$$L_U(F) = \int_{\mathcal{X}} q(x) \sum_{y \in \{1,-1\}} U(F(x)y - b(x; F))$$

$$- p(y|x)(F(x)y - b(x; F)) \, dx.$$  \hspace{1cm} (3.14)

For the constant volume condition and the normalized condition, the function $b(x; F)$ is represented as $b(F(x))$, since $b(x; F) = b(x; -F)$ holds and $b$ depends on $x$ through $F(x)$. Let us consider a function

$$V(z) = U(-z-b(z)) + U(z-b(z)) + z + b(z).$$

Then the above loss function, equation 3.14, is rewritten as $V$-loss function as follows:

$$L_V(F) = \int_{\mathcal{X}} q(x)\{p(1|x)V(-F(x)) + p(-1|x)V(F(x))\} \, dx.$$  \hspace{1cm} (3.15)

For a fixed $x$, we focus on a point-wise property of the $V$-loss and denote $p(1|x)$ by $\xi$ and $F(x)$ by $\gamma$. Let us define

$$C_{\xi}(\gamma) = \xi V(-\gamma) + (1 - \xi)V(\gamma),$$  \hspace{1cm} (3.16)

for $0 \leq \xi \leq 1$. The large sample behavior of the algorithm based on the above $V$-loss function was investigated.

**Definition 1.** The $V$-loss is classification calibrated if, for any $\xi \neq \frac{1}{2}$,

$$\inf_{\gamma: \gamma(2\xi-1) \leq 0} C_{\xi}(\gamma) > \inf_{\gamma \in \mathbb{R}} C_{\xi}(\gamma).$$  \hspace{1cm} (3.17)

The classification calibration implies that a minimizer of $C_{\xi}(\gamma)$ with respect to $\gamma \in \mathbb{R}$ has the same sign as $\xi - \frac{1}{2}$. If the loss function $V$ is classification calibrated, asymptotic convergence to the optimal value of loss implies that the decision function reaches the Bayes optimal decision rule. The following theorem is shown in Bartlett, Jordan, and McAuliffe (2006).

**Theorem 3.** The following conditions are equivalent:

1. $V$-loss is classification calibrated.
2. For every sequence of measurable functions $F_i : \mathcal{X} \to \mathbb{R}$, $i = 1, 2, \ldots$, and every probability distribution $q(x)p(y|x)$ on $\mathcal{X} \times \{1, -1\}$, 

$$\lim_{i \to \infty} \int_{\mathcal{X}} q(x) \sum_{y \in \mathcal{Y}} p(y|x)V(-yF_i(x)) \, dx$$

$$= \int_{\mathcal{X}} q(x) \sum_{y \in \mathcal{Y}} p(y|x)V(-yF_{bayes}(x)) \, dx$$

implies

$$\lim_{i \to \infty} \int_{\mathcal{X}} q(x) \sum_{y \in \mathcal{Y}} p(y|x)I(Y \neq sgn(F_i(x))) \, dx$$

$$= \int_{\mathcal{X}} q(x) \sum_{y \in \mathcal{Y}} p(y|x)I(Y \neq sgn(F_{bayes}(x))) \, dx.$$
Then we observe that the first and second derivatives of $V^v$ with respect to $\gamma$ are

$$V^v(\gamma) = \frac{2u(\gamma - b^v)}{u(-\gamma - b^v) + u(\gamma - b^v)}.$$  

$$V''^v(\gamma) = \frac{4u(-\gamma - b^v)^2u'(\gamma - b^v) + u(\gamma - b^v)^2u'(-\gamma - b^v)}{(u(-\gamma - b^v) + u(\gamma - b^v))^3}.$$  

Now we assume that function $U$ is a convex function with a positive derivative. Consequently we observe $V''^v(\gamma) > 0$, which implies a convexity of $V^v(\gamma)$ and $V''^v(0) = 1 > 0$. Then the $U$-loss under the constant volume condition is classification calibrated.

**Lemma 3.** Under the normalized condition, the $U$-loss is classification calibrated.

**Proof.** The function $V$ is written as $V^v(\gamma) = U(-\gamma - b^v) + U(\gamma - b^v) + \gamma + b^n$, where $b^n$ is a bias function satisfying the normalized condition 3.10. Using the condition, we obtain 

$$u(\gamma - b^n) + u(-\gamma - b^n) = 1.$$  

The first and second derivative of $V^n$ are calculated as

$$V''^n(\gamma) = u(-\gamma - b^n)(-1 - b^n) + u(\gamma - b^n)(1 - b^n) + 1 + b''^n,$$

$$V''''^n(\gamma) = u'(-\gamma - b^n)(-1 - b^n)^2 + u'(\gamma - b^n)(1 - b^n)^2.$$  

From those equations, we observe 

$$V''^n(0) > 0, \quad V''''^n(\gamma) > 0,$$  

which concludes that $U$-loss is classification calibrated under the normalized condition.

For the moment matching condition, the form of loss function differs from the one of $V$-loss (see equation 3.15) because the bias function is defined using the posterior distribution $p(y|x)$. Then $U$-loss under the moment matching condition is definitely not classification calibrated. However, we remark the following property, which seems to be essential for results in Bartlett et al. (2006):

**Remark 2.** Under the moment matching condition, the $U$-loss satisfies equation 3.17.
For the multiclass problem, the concept of the classification calibrated was proposed (Tewari & Bartlett, 2007). In future work, $U$-losses for multiclass problems will be classification calibrated.

For the special form of function $U_\eta(z) = (1 - \eta) \exp(z) + \eta z$, robust methods in the binary labels setting have been proposed under both the normalized and the moment matching conditions. Model 2.6 with the constant mislabeling probability, which is associated with the normalized condition, was statistically analyzed in Copas (1988). The $\eta$-Boost for binary classification problems in Takenouchi and Eguchi (2004), a modified version of AdaBoost for purposes of robustness, is introduced from the moment matching condition and is associated with a probabilistic model of $x$-dependent mislabeling. The associative model suggests that the larger the proportion of mislabeling, the nearer the feature vector is to the decision boundary. This model is useful and practical in many applications of pattern recognition. A major objective of this letter is to extend this $\eta$-Boost for binary classification to that for multiclass classification problems, and we recognize failure of the direct extension of binary-class $\eta$-Boost as based on the moment matching condition. Therefore, we present a new idea for an associated model of mislabeling in multiclass situations based on the constant volume condition 3.5. In the following section, we primarily focus on the constant volume condition 3.5, which leads to a mislabeling model 3.3 suitable for the multiclass problem. Before that, however, we consider the normalized condition, 3.10, that leads to the constant mislabeling probability model by way of introduction.

4 Eta-Divergence and Eta-Loss

Let us focus on the specific form of the function $U_\eta(z) = (1 - \eta) \exp(z) + \eta z (0 \leq \eta < 1)$ in the class of the Bregman $U$-divergences. By applying the argument in section 3 to $U_\eta$, we observe that Eta-divergence is

$$D_{U_\eta}(m, \mu) = \int_X q(x) \sum_{y \in Y} \left\{ \mu(x, y) - m(x, y) 
+ (m(x, y) - \eta) \log \frac{m(x, y) - \eta}{\mu(x, y) - \eta} \right\} dx,$$

and so the corresponding loss functions are

$$L_\eta(H) = \int_X q(x) \sum_{y \in Y} \left\{ (1 - \eta) \exp(H(x, y) - b_H(x)) + \eta (H(x, y) - b_H(x)) 
- p(y|x)(H(x, y) - b_H(x)) \right\} dx.$$  \hspace{1cm} (4.2)

Fixing $\eta = 0$, we obtain the extended Kullback-Leibler divergence $D_{U_\eta}$ and observe the relationship $D_{U_\eta}(p, q) = D_{U_\eta}(p + \eta, q + \eta)$. 

As noted above, we first introduce a method associated with the normalized condition, which leads to a constant probability of mislabeling; then we discuss a method associated with the constant volume condition, which induces the model of mislabeling, equation 2.13.

4.1 The Normalized Eta-Boost.M: The Normalized Condition and the Associative Model. The normalized condition given in equation 3.10 constrains the pseudoconditional model in equation 3.3 to be the probabilistic model. From the normalized condition with the function \( u_\eta(z) = U'_\eta(z) = (1 - \eta) \exp(z) + \eta \), we obtain the following relation:

\[
1 = \sum_{y \in \mathcal{Y}} u_\eta(H(x, y) - b_H^n(x)) = (1 - \eta) \exp(-b_H^n(x)) \sum_{y \in \mathcal{Y}} \exp(H(x, y)) + \eta G.
\]

The bias function can be written as

\[
b_H^n(x) = \log \frac{1 - \eta}{1 - G \eta} + \log \sum_{y \in \mathcal{Y}} \exp(H(x, y)),
\]

and the loss function is written as

\[
L_H^n(\eta) = (1 - \eta G) \left( 1 + \log \frac{1 - \eta}{1 - \eta G} \right) - \int_{\mathcal{X}} q(x) \sum_{y \in \mathcal{Y}} (p(y|x) - \eta) \log p_{0,H}(y|x) \, dx.
\]

where \( p_{0,H}(y|x) \) is the logistic model constructed by the discriminant function \( H(x, y) \), defined by equation 2.5. If \( \eta \) is set to 0, the above loss function is a negative log likelihood of the logistic model, and the optimization of equation 4.4 leads to the logistic discriminant analysis. From the loss function 4.4, we can derive normalized Eta-Boost.M as shown in appendix A.

From lemma 1, we observe the following relationship between the true conditional distribution \( p(y|x) \) of data and the optimal discriminant function \( H^*(x, y) = \arg\min_{H} L_H^n(\eta) \):

\[
p(y|x) = u_\eta(H^*(x, y) - b_H^*(x)) = (1 - \eta) \frac{1 - \eta G}{(1 - \eta) \sum_{y' \in \mathcal{Y}} \exp(H^*(x, y'))} \exp(H^*(x, y)) + \eta
\]

\[
= (1 - \eta G)p_{0,H^*}(y|x) + \eta
\]

\[
= (1 - \eta(G - 1))p_{0,H^*}(y|x) + \eta \sum_{y' \neq y} p_{0,H^*}(y'|x).
\]
This model represents the mislabeling model with constant probability $\eta$ and is a multiclass version of Copas (1988). In the following section, we introduce a model of mislabeling that depends on the feature $x$, derived from the constant volume condition 3.5 with the form of function $U_\eta(z) = (1 - \eta) \exp(z) + \eta z$.

4.2 The Constant Volume Condition and the Associative Model.
Under the constant volume condition with the function $U_\eta(z)$, we obtain the following theorem about the explicit form of the bias function $b_{H}(x)$:

**Theorem 4.** When we apply the constant volume condition 3.5 to the function $U_\eta(z) = (1 - \eta) \exp(z) + \eta z$ with the typical choice of constant $C = GU_\eta(0)$, we obtain the explicit form of the bias function $b_{H}(x)$ by solving the differential equation (see equation 3.8):

$$b_{H}(x) = \begin{cases} 
\frac{1}{G} \log \frac{1}{G} + \log \sum_{y' \in Y} \exp(H(x, y')), & \eta = 0, \\
\frac{1}{G} \sum_{y' \in Y} H(x, y') - \frac{1 - \eta}{\eta} + K(Q(H)), & 0 < \eta < 1,
\end{cases} \quad (4.6)$$

where $K(z)$ is the inverse function of $k(z) = z \exp(z)$ and

$$Q(H) = k \left( \frac{1 - \eta}{\eta} \right) \frac{1}{G} \sum_{y' \in Y} \exp \left( H(x, y') - \frac{1}{G} \sum_{y' \in Y} H(x, y') \right)$$

$$= k \left( \frac{1 - \eta}{\eta} \right) \exp(d_{Y}(p_{Ul}, p_{0,H}(\cdot|x))). \quad (4.7)$$

Here, $d_{Y}(\cdot, \cdot)$ is the Kullback-Leibler divergence on $Y \times Y$, defined by equation 2.11. The $p_{Ul}(y|x)$ is the uniform distribution on $Y$ given $x$ and $p_{0,H}$ is the logistic model 2.5 constructed by $H$.

**Proof.** In the following, we denote $H(x, y)$ by $H_{y}$ and $b_{H}^{y}(x)$ by $b$ when there is no risk of causing confusion.

First, we discuss the case where $\eta = 0$. From the constant volume condition 3.5, we immediately obtain the relation

$$\sum_{y \in Y} \exp(H_{y} - b) = C = G \exp(0). \quad (4.8)$$
The bias function with \( \eta = 0 \) is thus written as

\[
  b = \log \frac{1}{G} + \log \sum_{y \in \mathcal{Y}} \exp(H_y).
\]

Second, we proceed to the case where \( 0 < \eta < 1 \). By setting

\[
  r = \sum_{y \in \mathcal{Y}} \exp(H_y - b),
\]

we obtain

\[
  \frac{\partial r}{\partial H_y} = \exp(H_y - b) - \frac{\partial b}{\partial H_y} \sum_{y' \in \mathcal{Y}} \exp(H_{y'} - b)
  = r \left( p_{0,H}(y|x) - \frac{\partial b}{\partial H_y} \right),
\]

\[
  \frac{\partial b}{\partial H_y} = \frac{(1 - \eta) \exp(H_y - b) + \eta}{\sum_{y' \in \mathcal{Y}} ((1 - \eta) \exp(H_{y'} - b) + \eta)}
  = \frac{(1 - \eta) r p_{0,H}(y|x) + \eta}{(1 - \eta) r + \eta G}.
\]

From the first line to the second line, we use equation 3.8 with the form \( u = u_\eta \). From the above two equations, we observe that

\[
  \frac{\partial r}{\partial H_y} = \frac{r \eta G}{(1 - \eta) r + \eta G} \left\{ p_{0,H}(y|x) - \frac{1}{G} \right\},
\]

(4.10)

It is possible to solve equation 4.10 by the separation of variables:

\[
  \int \left( \frac{1 - \eta}{\eta G} + \frac{1}{r} \right) dr = \int \left( p_{0,H}(y|x) - \frac{1}{G} \right) dH_y,
\]

(4.11)

\[
  \frac{(1 - \eta) r}{\eta G} + \log r = \log \sum_{y' \in \mathcal{Y}} \exp(H_{y'}) - \frac{H_y}{G} + C,
\]

(4.12)

where \( C \) is a function of \( H_{y'} \), \( y' \neq y \). Since we can use the same way to all \( y \in \mathcal{Y} \), the following equation is obtained:

\[
  \frac{(1 - \eta) r}{\eta G} + \log r = \log \sum_{y' \in \mathcal{Y}} \exp(H_{y'}) - \frac{1}{G} \sum_{y' \in \mathcal{Y}} H_{y'} + C',
\]

(4.13)
where $C'$ is a constant term that does not depend on $H_1, \ldots, H_y$. If $C'$ is subject to the initial condition that $C = GU_η(0)$, which implies $b^v_H = 0$ when $H_1 = \cdots = H_G = 0$, we obtain $C' = \frac{1-η}{η}$. We can rewrite equation 4.13 as

$$
\frac{1 - η}{ηG} \exp \left( \frac{1 - η}{ηG} r \right) = k \left( \frac{1 - η}{ηG} r \right) = \frac{1 - η}{ηG} \exp \left( \frac{1 - η}{η} \right) \sum_{y' \in Y} \exp \left( H_{y'} - \frac{1}{G} \sum_{y'' \in Y} H_{y''} \right) = Q(H).
$$

(4.14)

From the above equation and the relation $K = k^{-1}$, we obtain

$$
b = \log \frac{1-η}{ηG} + \log \sum_{y' \in Y} \exp(H_{y'}) - \log K(Q(H)).
$$

(4.15)

Additionally, we observe that

$$
\log z = \log k(K(z)) = \log (K(z) \exp(K(z))) = K(z) + \log K(z).
$$

(4.16)

By applying equation 4.16 to equation 4.13, we obtain

$$
b = \log \frac{1-η}{ηG} + \log \sum_{y' \in Y} \exp(H_{y'}) + K(Q(H)) - \log Q(H) = \frac{1}{G} \sum_{y' \in Y} H_{y'} - \frac{1-η}{η} + K(Q(H)).
$$

(4.17)

Thus, we conclude the bias function, equation 4.6.

Note that since $Q(H)$ takes a positive value for any discriminant function $H(x, y)$, $K(z)$ is the positive and monotonically increasing function in our context. The loss function with the bias function 4.6 has the form of equation 3.9 and we observe that the loss function with $η = 0$, $L^v_0(H)$, also coincides with the negative log likelihood of the logistic model 2.5, as in the case of the loss function 4.4 with $η = 0$ except for a constant term. The algorithm Eta-Boost.M is derived from the sequential minimization of the loss function $L^v_η(H)$.

We discuss a relation between $L^v_η(H)$, the eta-loss function 4.2 with the constant volume condition and the optimal discriminant function $H^*(x, y)$. 

From theorem 2, we observe

\[ p(y|x) = \frac{(1 - \eta) \exp \left( H^*(x, y) - b_{H^*}^T(x) \right) + \eta \sum_{y' \in Y} (1 - \eta) \exp \left( H^*(x, y') - b_{H^*}^T(x) \right) + \eta G}{\sum_{y' \in Y} (1 - \eta) \exp \left( H^*(x, y') - b_{H^*}^T(x) \right) + \eta G} \]

\[ = (1 - (G - 1)\delta(H^*)) p_{0, H^*}(y|x) + \delta(H^*) \sum_{y' \neq y} p_{0, H^*}(y'|x). \quad (4.18) \]

where \( \delta(H) \) can be interpreted as a probability of mislabeling:

\[ \delta(H) = \frac{\eta}{\sum_{y' \in Y} (1 - \eta) \exp \left( H(x, y') - b_{H}^T(x) \right) + G \eta} = \frac{1}{G \left[ 1 + K \left( k \left( 1 - \eta \right) \exp \left( d_Y(p_U, p_{0, H}(\cdot|x)) \right) \right) \right]}. \quad (4.19) \]

The \( \delta(H^*) \) represents greater association depending on \( x \) in the following sense. From the property of the KL divergence, we obtain

\[ \delta(H^*) \leq \frac{1}{G \left[ 1 + K \left( k \left( 1 - \eta \right) \right) \right]} = \frac{\eta}{G}, \quad (4.20) \]

and the equality holds if and only if \( p_{0, H^*} = p_U \). The optimal discriminant function \( H^*(x, y) \) has no information about a class label of the feature \( x \). The probability of mislabeling, \( \delta(H) \), rapidly decreases as the feature vector \( x \) diverges from the common decision boundary because the function \( K(\cdot) \) with a nonnegative argument is a monotonically increasing function. Figure 1 shows \( \delta(H) \) against \( \eta \) and KL divergence, \( d_Y(p_U, p_{H, 0}(\cdot|x)) \) at \( G = 10 \). Eta-Boost.M provides a mislabeling mechanism such that a smaller proportion is given in accordance with the difficulty of classification of the input \( x \). In other words, Eta-Boost.M reduces the risk of a hasty decision for a region where it is difficult to discriminate.

**4.3 Comparison of Eta-Boost.M with the Binary \( \eta \)-Boost.** For the binary classification problem with a label \( y \in \{1, -1\} \), the binary \( \eta \)-Boost has been proposed by considering the model of mislabeling associated with the input \( x \) in Takenouchi and Eguchi (2004). In this section, we compare Eta-Boost.M with the binary \( \eta \)-Boost algorithm. The binary \( \eta \)-Boost algorithm includes the original AdaBoost as a special case and is derived from the sequential minimization of the mixture of the exponential loss function and the naive error loss function. Although a formal extension of
binary $\eta$-Boost to a multiclass classification problem is possible, the statistical properties related to the mislabeling are unclear, as in the previous section. Consequently, we focus on the degree of deviation from the uniform distribution as a criterion for classification difficulty and technically leap from the binary $\eta$-Boost to Eta-Boost.M by introducing the constant volume condition.

In the binary $\eta$-Boost algorithm, the discriminant function $H$ is defined by $H(x, y) = F(x)y$, where $F(x)$ is a binary classifier, and a probability of mislabeling is modeled as the following form:

$$
\delta'(H) = \frac{1}{2\left\{ 1 + \frac{1-\eta}{\eta} \exp(d_Y(p_U, p_{H,0}(\cdot|x))) \right\}}.
$$

(4.21)

Then the mislabeling probability 4.19 coincides with 4.21 if we substitute $k(z) = z \exp(z)$ for the identity function $z$. If the quantity $d_Y(p_U, p_{H,0}(\cdot|x))$ is sufficiently small, we can approximate equations 4.19 and 4.21 as

$$
\delta(H) \approx \frac{\eta}{G} - \frac{\eta^2(1-\eta)}{G}d_Y(p_U, p_{H,0}(\cdot|x))
$$

(4.22)
Figure 2: The probability of mislabeling of two methods against $\eta$. The quantity $d_Y$ is fixed to 0.5.

and

$$\delta'(H) \approx \frac{\eta}{G} - \frac{\eta(1-\eta)}{G}d_Y(p_U, p_{H,0}(-|x)).$$

(4.23)

Those equations imply that Eta-Boost.M assumes a higher probability of mislabeling compared with binary $\eta$-Boost and tends to make a conservative decision about class labels. Figure 2 shows the probability of mislabeling of each method against $\eta$ when we fix the KL divergence $d_Y$ to 0.5.

4.4 Weight Property of Eta-Boost.M. The general algorithm derived from the $U$-loss function $L_U(H)$ is described in appendix A, and the error rate property of the algorithm under the constant volume condition is shown in appendix B. In this section, we discuss the property of the weight associated with the function $U_\eta$ under the normalized condition and the constant volume condition. From equations B.4 and 4.5 or 4.18, the weight function under the normalized condition, $D^\eta_t(i, y)$, and one under the constant volume condition, $D^\eta_v(i, y)$, are written as

$$D^\eta_t(i, y) = \frac{1}{N} \left( (1-\eta(G-1))p_{0,H-1}(y|x) + \eta \sum_{y' \neq y} p_{0,H-1}(y'|x) \right)$$

(4.24)
and

\[
D^*_i(i, y) = \frac{1}{N} \left( (1 - \delta(H_{t-1})(G - 1))p_{0,H_{t-1}}(y|x) \right.
\]
\[
+ \delta(H_{t-1}) \sum_{y' \neq y} p_{0,H_{t-1}}(y'|x) \right),
\]

(4.25)

While the weight function 4.24 assumes the constant probability of mislabeling throughout the algorithm’s process, equation 4.25 depends on the probability of mislabel \(\delta(H_{t-1})\) when the learning process proceeds because \(d_Y(p_U, p_{0,H_{t-1}}(\cdot|x))\) tends to take larger values as the algorithm progresses. This implies that the weight function is regularized to avoid excessively premature discrimination.

5 Simulation Studies

In this section, we investigate the performance of the proposed methods using artificial data sets in the context of the multiclass classification. Also, we apply methods to UCI repository data sets (Blake & Merz, 1998). For comparison, we employ several methods, namely AdaBoost.M2; the AdaBoost.M1W-type algorithm (Eibl & Pfeiffer, 2002), which is a GrPloss method (Eibl & Pfeiffer, 2005), with decision stumps making hard decisions; linear discriminant analysis, quadratic discriminant analysis; and rpart, a decision tree in the R (R Development Core Team, 2006). The weak hypothesis, 2.4, is constructed by the decision stump,

\[
\hat{y}(x) = \begin{cases} 
  S, & x_j \geq b, \\
  \mathcal{Y} \setminus S, & \text{otherwise},
\end{cases}
\]

where \(S\) is a subset of \(\mathcal{Y}\), \(x_j\) is the \(j\)th variable of \(x\), and \(b\) is a threshold value. In our experiments, we considered all patterns as the subset \(S\).

5.1 Synthetic Data Set. We examined a case in which a training data set was generated by model 4.18, and the optimal discriminant function is

\[
H^*(x, y) = \begin{cases} 
  x_1, & y = 1, \\
  -x_1, & y = 2, \\
  20 - \frac{x_1^2 + x_2^2}{3.2}, & y = 3.
\end{cases}
\]

(5.1)

Let \(q(x)\) be the probability density function of the two-dimensional uniform distribution on \((-10, 10) \times (-10, 10)\), and we fix \(\eta = 0.5\) in the conditional
probability of the class label, equation 4.18. We generated 50 different training data sets and test data sets and investigated the averaged performance of Eta-Boost.M. Tuning parameters $\eta$, $T$ were determined by five-fold cross-validation for each training data set. The test data set contains 3000 observations. A typical training data set containing 300 observations and the classification boundary of the Bayes rule is shown in Figure 3.

Contaminated examples were observed particularly near the boundary. The probability of mislabeling, equation 4.19, associated with equation 5.1 is shown in Figure 4, and we can confirm that the probability has the maximum value at the intersection of boundaries of all classes.

Figure 5 shows the averaged test errors of Eta-Boost.M, normalized Eta-Boost.M, AdaBoost.M2, and AdaBoost.M1W method in 50 different data sets against the learning step $t$. We observed that Eta-Boost.M at the learning step $T$.Eta, performed well and attained lower error compared to AdaBoost.M2 at the learning steps $T$.Ada.

Figure 6 shows the averaged test error and cross-validation error of Eta-Boost.M and the test error of normalized Eta-Boost.M against $\eta$. Eta-Boost.M outperformed AdaBoost.M2, AdaBoost.M1W, linear discriminant analysis (0.32339), quadratic discriminant analysis (0.22271), and rpart (0.19827) in
Figure 4: The mislabeling probability, equation 4.19, associated with equation 5.1.

Figure 5: The average of test errors of Eta-Boost.M ($\eta = 0.2$), AdaBoost.M2, normalized Eta-Boost.M ($\eta = 0.05$), and AdaBoost.M1W method.
Figure 6: The average of test errors of Eta-Boost.M and normalized Eta-Boost.M and cross-validation errors of Eta-Boost.M are plotted against $\eta$. Additionally, averaged test errors of AdaBoost.M2, QDF, rpart, and AdaBoost.M1W are plotted.

terms of test error for appropriate $\eta$, and we could determine the appropriate $\eta$ and $T$ by cross-validation error. The error of normalized Eta-Boost.M tended to diverge as $\eta$ increased, and we found that appropriate tuning of $\eta$ was difficult for this data set. Additionally, we applied a multiclass version of SVM (Crammer & Singer, 2001) with a polynomial kernel (0.31471) and a radial basis function (RBF) kernel (0.14819). The SVM with the RBF kernel outperformed Eta-Boost.M; however, its performance is explained by the nonlinearity of the kernel function and depends on a selection of kernel function. Our focus is not on the effect of the kernel function but on the method for combining weak learning machines.

Figure 6 shows only the averaged improvement of the proposed methods. Then, we have examined the superiority of the proposed method to the existing method for each data set. For this purpose, we have investigated the winning rate of Eta-Boost.M in 50 experiments against other methods, shown in Figure 7. Consequently we see that Eta-Boost.M with an appropriate $\eta$ outperforms other methods with a high probability in terms of test error.
Figure 7: Winning rate of Eta-Boost.M in 50 runs for other methods in the sense of test error.

Figures 8 and 9, respectively, show boundaries of AdaBoost.M2 and Eta-Boost.M ($\eta = 0.2$). The optimal learning step for each method is estimated by cross-validation. While the boundary of AdaBoost.M2 is influenced by examples that are difficult to classify, Eta-Boost.M tends to construct a comparatively smooth boundary.

5.2 UCI Repository. We next applied the proposed methods, Eta-Boost.M and normalized Eta-Boost.M, to data sets in the UCI repository. Each data set was originally divided into a training data set and a test data set except for the waveform data set. The waveform data set was randomly divided into a training data set and a test data set. More detailed information on the data sets is given in Table 1.

First, we applied the proposed methods, AdaBoost.M1W method and AdaBoost.M2, for the data set. In all applied methods, the number $T$ of the learning step should be determined, and Eta-Boost.M and normalized
Figure 8: A typical training data set, a classification boundary of Bayes’ rule, and a classification boundary of AdaBoost.M2.

Figure 9: A typical training data set, a classification boundary of Bayes’ rule, and a classification boundary of Eta-Boost.M, $\eta = 0.2$. 
Table 1: Information on Data Sets.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Number of Total</th>
<th>Number of Training</th>
<th>Number of Test</th>
<th>Number of Attributes</th>
<th>Number of Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image</td>
<td>2310</td>
<td>210</td>
<td>2100</td>
<td>19</td>
<td>7</td>
</tr>
<tr>
<td>Satellite</td>
<td>6435</td>
<td>4435</td>
<td>2000</td>
<td>36</td>
<td>6</td>
</tr>
<tr>
<td>Pendigit</td>
<td>10,992</td>
<td>7494</td>
<td>3498</td>
<td>16</td>
<td>10</td>
</tr>
<tr>
<td>Waveform</td>
<td>5000</td>
<td>3000</td>
<td>2000</td>
<td>21</td>
<td>3</td>
</tr>
<tr>
<td>Thyroid</td>
<td>7200</td>
<td>3772</td>
<td>3428</td>
<td>21</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 2: Training and Test Error Rate for Data Sets in UCI Repository and for Contaminated Data Sets.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Error</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.10476</td>
</tr>
<tr>
<td>Satellite</td>
<td>0.09290</td>
<td>0.07734</td>
<td>0.04397</td>
<td>0.18625</td>
</tr>
<tr>
<td>Pendigit</td>
<td>0.01468</td>
<td>0.01161</td>
<td>0.00961</td>
<td>0.21390</td>
</tr>
<tr>
<td>Waveform</td>
<td>0.13033</td>
<td>0.13333</td>
<td>0.11267</td>
<td>0.15367</td>
</tr>
<tr>
<td>Waveform (10%)</td>
<td>0.21300</td>
<td>0.19933</td>
<td>0.21967</td>
<td>0.22333</td>
</tr>
<tr>
<td>Waveform (20%)</td>
<td>0.29133</td>
<td>0.28967</td>
<td>0.28700</td>
<td>0.30200</td>
</tr>
<tr>
<td>Thyroid</td>
<td>0.00080</td>
<td>0.00000</td>
<td>0.00053</td>
<td>0.02174</td>
</tr>
<tr>
<td>Thyroid (10%)</td>
<td>0.11612</td>
<td>0.11532</td>
<td>0.11320</td>
<td>0.12540</td>
</tr>
<tr>
<td>Thyroid (20%)</td>
<td>0.22004</td>
<td>0.21633</td>
<td>0.21739</td>
<td>0.22423</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Error</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image</td>
<td>0.06476</td>
<td>0.05571</td>
<td>0.06476</td>
<td>0.12810</td>
</tr>
<tr>
<td>Satellite</td>
<td>0.13300</td>
<td>0.12950</td>
<td>0.11650</td>
<td>0.19250</td>
</tr>
<tr>
<td>Pendigit</td>
<td>0.05975</td>
<td>0.06089</td>
<td>0.05746</td>
<td>0.24271</td>
</tr>
<tr>
<td>Waveform</td>
<td>0.14550</td>
<td>0.15800</td>
<td>0.14650</td>
<td>0.15100</td>
</tr>
<tr>
<td>Waveform (10%)</td>
<td>0.17000</td>
<td>0.16900</td>
<td>0.16750</td>
<td>0.16450</td>
</tr>
<tr>
<td>Waveform (20%)</td>
<td>0.17650</td>
<td>0.15750</td>
<td>0.15750</td>
<td>0.17800</td>
</tr>
<tr>
<td>Thyroid</td>
<td>0.00729</td>
<td>0.00758</td>
<td>0.00642</td>
<td>0.03034</td>
</tr>
<tr>
<td>Thyroid (10%)</td>
<td>0.02917</td>
<td>0.02246</td>
<td>0.02100</td>
<td>0.03792</td>
</tr>
<tr>
<td>Thyroid (20%)</td>
<td>0.02946</td>
<td>0.02480</td>
<td>0.02392</td>
<td>0.03151</td>
</tr>
</tbody>
</table>

Eta-Boost.M additionally had a hyperparameter $\eta$ that should be appropriately tuned. Those parameters $T$, $\eta$ were determined by five-fold cross-validation using only the training examples, as in the case of the experiment with synthetic data sets. Second, we investigated the performance of the proposed methods using the data sets Thyroid and Waveform with simulating a random mislabeling in the label observation process, where the class label of the data set was randomly contaminated with a constant probability. We examined two cases of the contamination probability, 0.1 and 0.2. Note that this contamination process was different from that assumed by the statistical model, equation 4.18, associated with the constant volume condition.
Training and test performances for the original data set and those two contaminated data sets are shown in Table 2, and the best performance for each data set is in bold type. For contaminated data sets, normalized Eta-Boost.M outperforms the other methods because the mechanism of mislabeling coincides with the assumed one. Additionally, Eta-Boost.M worked well even though the assumed mechanism of mislabeling is different from the optimal one, equation 4.18.

6 Conclusion

We proposed a new class of loss function associated with the constant volume condition in the boosting algorithm derived from the minimization of the Bregman $U$-divergence. As a main result, we derived two kinds of robust boosting algorithms, normalized Eta-Boost.M and Eta-Boost.M, which naturally associate with mechanisms of mislabeling. The former algorithm considers the uniform contamination among multiclass labels, while the latter makes it possible to consider the contamination depending on the input. The method assumes a larger proportion of mislabeling for a region of input space where it is difficult to discriminate. This mechanism reduces the risk of a hasty decision.

We have numerically explored the robustness of two versions of Eta-Boost.M in regard to the appearance of mislabeling with the synthetic data set. Selection of the tuning parameter $\eta$ in the proposed methods was implemented by the $K$-fold cross-validation error rate, which reinforces the advantage of the proposed methods over AdaBoost.M2 or other existing methods. Also for data sets in the UCI repository, we investigated the performance of the proposed methods and confirmed that the proposed methods attained better performance on robustness for noisy data sets. Theoretical analysis of the large sample behavior of the algorithm for multiclass problem remains to be investigated.

Appendix A: A Boosting Algorithm Associated with $U$-Loss

Here we derive a boosting algorithm associated with the $U$-loss. For a given data set $\{(x_i, y_i) : i = 1, \ldots, N\}$, the empirical version of $U$-loss, equation 3.4, is written as

$$\bar{L}_U(H) = \frac{1}{N} \sum_{i=1}^{N} \left\{ -H(x_i, y_i) + b_H(x_i) + \sum_{y \in Y} U(H(x_i, y) - b_H(x_i)) \right\}.$$ 

We can obtain the algorithm by the sequential minimization of the above loss function. It is assumed that a discriminant function $H_t(x, y)$ is obtained and we consider an update from $H_t$ to $H_t + \alpha h$, where $h$ is a new weak
hypothesis in $\mathcal{H}$ and $\alpha$ is a small constant. We then observe that
\[
\bar{L}_U(H_t + \alpha h) - \bar{L}_U(H_t)
\approx \frac{\alpha}{N} \sum_{i=1}^{N} \sum_{y \in \mathcal{Y}} \left\{ u(H_t(x_i, y) - b_{H_t}(x_i))(h(x_i, y) - \frac{\partial b_{H_t + \alpha h}(x_i)}{\partial \alpha} \bigg|_{\alpha=0}) \right\}
- \left( h(x_i, y) - \frac{\partial b_{H_t + \alpha h}(x_i)}{\partial \alpha} \bigg|_{\alpha=0} \right)
= \frac{\alpha}{N} \sum_{i=1}^{N} \sum_{y \in \mathcal{Y}} \left\{ u(H_t(x_i, y) - b_{H_t}(x_i))
- \bar{p}(y|x_i) \right\} \left\{ h(x_i, y) - \frac{\partial b_{H_t + \alpha h}(x_i)}{\partial \alpha} \bigg|_{\alpha=0} \right\},
\] (A.1)
where $\bar{p}(y|x)$ is an empirical version of the conditional distribution $p(y|x)$. The quantity A.1 is a functional of the hypothesis $h(x, y)$, and the optimization of $h$ does not depend on $\alpha$. We define $\mathcal{E}_t(h)$ by
\[
\mathcal{E}_t(h) = \frac{1}{N} \sum_{i=1}^{N} \sum_{y \in \mathcal{Y}} \left\{ u(H_t(x_i, y) - b_{H_t}(x_i))
- \bar{p}(y|x_i) \right\} \left\{ h(x_i, y) - \frac{\partial b_{H_t + \alpha h}(x_i)}{\partial \alpha} \bigg|_{\alpha=0} \right\},
\] (A.2)
and select a new hypothesis $h_{t+1}(x, y)$ by minimizing $\mathcal{E}_t(h)$ as
\[
h_{t+1} = \arg\min_{h \in \mathcal{H}} \mathcal{E}_t(h).
\]
For the selected hypothesis $h_{t+1}(x, y)$, we optimize the coefficient $\alpha_{t+1}$ by the line search,
\[
\alpha_{t+1} = \arg\min_{\alpha} \bar{L}_U(H_t + \alpha h_{t+1}).
\]
Then we obtain an updated discriminant function $H_{t+1} = H_t + \alpha_{t+1} h_{t+1}$. A general algorithm of boosting is summarized as follows:

1. Setting $H_0(x, y) = 0$.
2. for $t = 1, \ldots, T$
   (a) Find $h_t = \arg\min_{h \in \mathcal{H}} \mathcal{E}_{t-1}(h)$.
   (b) Calculate $\alpha_t = \arg\min_{\alpha} \bar{L}_U(H_{t-1} + \alpha h_t)$.
   (c) Update the discriminant function as $H_t = H_{t-1} + \alpha h_t$.
3. Output the discriminant function $H_T = \sum_{t=1}^{T} \alpha_t h_t(x, y)$.
A label for a new feature $x$ is predicted by $\arg\max_y H_T(x, y)$. 
Appendix B: Error Rate Property of the Algorithm Under the Constant Volume Condition

Here we discuss criterion A.2 for selecting a new weak hypothesis based on the given data set and the associated empirical distributions \( \tilde{q}(x) \) and \( \tilde{p}(y|x) \). Under the normalized condition 3.10 and the moment matching condition 3.12, we pointed out that equation A.2 in the general algorithm can be interpreted as a weighted error rate (Murata et al., 2004). Also, under the constant volume condition 3.5, we can rewrite step 2a of the general algorithm as an optimization of a weighted error rate. By differentiating equation 3.5 with respect to \( \alpha \), we obtain

\[
0 = \frac{\partial}{\partial \alpha} \sum_{y \in Y} U(H(x, y) + \alpha h(x, y) - b_{H+\alpha h}(x)) \bigg|_{\alpha = \beta} \\
= \sum_{y \in Y} u(H(x, y) + \beta h(x, y) - b_{H+\beta h}(x)) \left( h(x, y) - \left. \frac{\partial b_{H+\alpha h}(x)}{\partial \alpha} \right|_{\alpha = \beta} \right).
\]

and then

\[
\left. \frac{\partial b_{H+\alpha h}(x)}{\partial \alpha} \right|_{\alpha = \beta} = \frac{\sum_{y \in Y} u(H(x, y') + \beta h(x, y') - b_{H+\beta h}(x)) h(x, y')}{\sum_{y' \in Y} u(H(x, y') + \beta h(x, y') - b_{H+\beta h}(x))}.
\]

By substituting equation B.2 with \( \beta = 0 \) into A.2, we observe that

\[
\mathcal{E}_t(h) = \frac{1}{N} \sum_{i=1}^{N} \sum_{y \in Y'} \tilde{p}(y|x_i) \left( -h(x_i, y) + \sum_{y' \in Y} u(H_t(x_i, y') - b_{H_t}(x_i)) h(x_i, y') \right) \\
= \frac{1}{N} \sum_{i=1}^{N} \left( \sum_{y' \in Y} u(H_t(x_i, y') - b_{H_t}(x_i)) h(x_i, y') - \left( \sum_{y' \in Y} u(H_t(x_i, y') - b_{H_t}(x_i)) \right) h(x_i, y_i) \right) \\
= \frac{1}{N} \sum_{i=1}^{N} \sum_{y \in Y'} \left( h(x_i, y) - h(x_i, y_i) \right) \frac{u(H_t(x_i, y) - b_{H_t}(x_i))}{\sum_{y' \in Y} u(H_t(x_i, y') - b_{H_t}(x_i))}.
\]

Let us define the weight for a pair comprising the \( i \)th example and the index \( y \) at step \( t \) by

\[
D_t(i, y) = \frac{1}{N} \frac{u(H_{t-1}(x_i, y) - b_{H_{t-1}}(x_i))}{\sum_{y' \in Y} u(H_{t-1}(x_i, y') - b_{H_{t-1}}(x_i))}
\]
and the weighted error rate of the hypothesis \( h \) by

\[
\varepsilon_t(h) = \sum_{i=1}^{n} \sum_{y \in Y} \frac{h(x_i, y) - h(x_i, y_i) + 1}{2} D_t(i, y). \tag{B.5}
\]

We discuss the factor of the right-hand side of equation B.5:

\[
\frac{h(x_i, y) - h(x_i, y_i) + 1}{2} = \begin{cases} 
0 & \text{if } y_i \in f(x_i) \text{ and } y \not\in f(x_i), \\
\frac{1}{2} & \text{if } y_i \in f(x_i) \text{ and } y \in f(x_i), \\
\frac{1}{2} & \text{if } y_i \not\in f(x_i) \text{ and } y \not\in f(x_i), \\
1 & \text{if } y_i \not\in f(x_i) \text{ and } y \in f(x_i),
\end{cases} \tag{B.6}
\]

where \( f(x) \) is a classification rule associated with the weak hypothesis \( h(x, y) \). If the classification machine \( f(x) \) cannot distinguish the label \( y_i \) from another label \( y \), it takes \( 1/2 \). On the other hand, it takes 0 when \( x_i \) is correctly classified as \( y_i \) and takes 1 when \( x_i \) is incorrectly classified in a situation where \( f(x_i) \) can distinguish the label \( y_i \) from the other label \( y \). We thus obtain the following relationship:

\[
\varepsilon_t(h) = 2\varepsilon_t(h) - 1.
\]

Consequently, step 2a in the general algorithm is equivalent to a minimization of the weighted error rate, equation B.5.

In Murata et al. (2004), the weighted error rate property \( \varepsilon_{t+1}(h_t) = 1/2 \) was observed under both the normalized condition and the moment matching condition. We obtain the same property for the weighted error rate B.5 as the following lemma.

**Lemma 4.** Under the constant volume condition 3.5, a relationship

\[
\varepsilon_{t+1}(h_t) = \frac{1}{2}, \tag{B.7}
\]

holds.

**Proof.** The result is immediately shown by the equilibrium condition of the coefficient \( \alpha_t \). Assuming that the discriminant function \( H_{t-1}(x, y) \) is obtained and the algorithm selects the new hypothesis \( h_t(x, y) \), then \( \alpha_t \) satisfies

\[
0 = \frac{\partial}{\partial \alpha} L^v_t(H_{t-1} + \alpha h_t) \bigg|_{\alpha = \alpha_t} = \frac{1}{N} \sum_{i=1}^{N} \left\{ -h_t(x_i, y_i) + \frac{\partial h_{t-1} + \alpha h_t(x_i)}{\partial \alpha} \bigg|_{\alpha = \alpha_t} \right\}
\]
\[ 
N \sum_{i=1}^{N} \sum_{y \in Y} (h_t(x_i, y_t) - h_t(x_i, y)) D_{t+1}(i, y) 
= 2\varepsilon_{t+1}(h_t) - 1. 
\]

(B.8)

Thus we conclude equation B.7.

This lemma shows that the selected \( t \)th hypothesis \( h_t \) can be regarded as the weakest hypothesis or as a random guess under the updated weight \( D_{t+1}(i, y) \) of the next \((t + 1)\)th step.

References


Received November 20, 2006; accepted July 15, 2007.
This article has been cited by: