Coherence Functions for Multicategory Margin-based Classification Methods

Zhihua Zhang
College of Comp. Sci. and Tech.
Zhejiang University
Zhejiang 310027, China
zhzhang@cs.zju.edu.cn

Michael I. Jordan
Depts. of EECS and Statistics
University of California, Berkeley
Berkeley, CA 94720, USA
jordan@cs.berkeley.edu

Wu-Jun Li and Dit-Yan Yeung
Dept. of Comp. Sci. and Eng.
Hong Kong Univ. of Sci. and Tech.
Hong Kong, China
{liwujun, dyyeung}@cse.ust.hk

Abstract

Margin-based classification methods are typically devised based on a majorization-minimization procedure, which approximately solves an otherwise intractable minimization problem defined with the 0-l loss. The extension of such methods from the binary classification setting to the more general multicategory setting turns out to be non-trivial. In this paper, our focus is to devise margin-based classification methods that can be seamlessly applied to both settings, with the binary setting simply as a special case. In particular, we propose a new majorization loss function that we call the *coherence function*, and then devise a new multicategory margin-based boosting algorithm based on the coherence function. Analogous to deterministic annealing, the coherence function is characterized by a temperature factor. It is closely related to the multinomial log-likelihood function and its limit at zero temperature corresponds to a multicategory hinge loss function.

1 Introduction

Margin-based classification methods have become increasingly popular since the advent of the support vector machine (SVM) (Cortes and Vapnik, 1995) and boosting (Freund, 1995; Freund and Schapire, 1997). These algorithms were originally designed for binary classification problems. Unfortunately, extension of them to the multicategory setting has been found to be non-trivial.

Appearing in Proceedings of the 12^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5.

A variety of *ad hoc* extensions of the binary SVM and boosting to multicategory classification problems have been studied. These include one-versus-rest, one-versus-one, error-correcting codes, and pairwise coupling (Allwein et al., 2000). Among these methods, one-versus-rest has been the dominant approach. The basic idea is to train m binary classifiers for an m-class ($m \geq 2$) problem so that each classifier learns to discriminate one class from the rest. However, optimality achieved for each of the m independent binary problems does not readily guarantee optimality for the original m-class problem.

The goal of this paper is to solve multicategory classification problems using the same margin principle as that for binary problems. Of crucial concern are the statistical properties (Bartlett et al., 2006; Tewari and Bartlett, 2007; Zhang, 2004) of a majorization function for the original 0-1 loss function. In particular, we analyze the Fisher-consistency properties (Zou et al., 2008) of extant majorization functions, which are built on the exponential, logit and hinge functions. This analysis inspires us to propose a new majorization function, which we call the *coherence function*.

The coherence function is attractive because it is a Fisher-consistent majorization of the 0-1 loss. Also, one limiting version of it is just the multicategory hinge loss function of Crammer and Singer (2001), and its relationship with the multinomial log-likelihood function is very clear. Moreover, this function is differentiable and convex. Thus it is very appropriate for use in the development of multicategory margin-based classification methods, especially boosting algorithms. Friedman et al. (2000) and Zou et al. (2008) proposed the multicategory LogitBoost and GentleBoost algorithms based on the multinomial log-likelihood function and the exponential loss function, respectively. We propose in this paper a new multicategory GentleBoost algorithm based on our coherence function.

The rest of this paper is organized as follows. Section 2 presents theoretical discussions of extant loss

functions for multicategory margin-based classification methods. Section 3 proposes the coherence function and discusses its statistical properties. Section 4 devises a multicategory margin-based boosting algorithm using the coherence function. An experimental analysis is presented in Section 5 and concluding remarks are given in Section 6. Some proofs of the theoretical results are left to the appendices.

2 Problem Formulation

We are given an m-class ($m \geq 2$) classification problem with a set of training data $\{(\mathbf{x}_i, c_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ is an input vector and $c_i \in \{1, 2, \ldots, m\}$ is its corresponding class label. We assume that each $\mathbf{x}$ belongs to one and only one class. Our goal is to find a classifier $\phi(\mathbf{x}) : \mathbf{x} \to c \in \{1, \ldots, m\}$.

Let $P_c(\mathbf{x}) = P(C = c|X = \mathbf{x})$, $c = 1, \ldots, m$ be the class probabilities given $\mathbf{x}$. The expected error at $\mathbf{x}$ is then defined by $\sum_{c=1}^m \mathbb{I}_{[\phi(\mathbf{x})\neq c]} P_c(\mathbf{x})$, where $\mathbb{I}_{[\#]} = 1$ if $\#$ is true and 0 otherwise. The empirical error on the training data is given by

$$\epsilon = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{[\phi(\mathbf{x}_i)\neq c_i]}.$$

Since ϵ is equal to its minimum value of zero when all the training data points are correctly classified, we wish to use ϵ as a basis for devising multicategory classification algorithms.

2.1 Multicategory Margins

Suppose the classifier is modeled using an m-vector $\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), \ldots, g_m(\mathbf{x}))^T$, where the induced classifier is obtained via maximization in a manner akin to discriminant analysis: $\phi(\mathbf{x}) = \text{argmax}_j \{g_j(\mathbf{x})\}$. For simplicity of analysis, we assume that the maximizing argument of $\max_j g_j(\mathbf{x})$ is unique. Of course this does not imply that the maximum value is unique; indeed, adding a constant to each component $g_j(\mathbf{x})$ does not change the maximizing argument. To remove this redundancy, it is convenient to impose a sum-to-zero constraint. Thus we define

$$\mathcal{G} = \left\{ (g_1(\mathbf{x}), \ldots, g_m(\mathbf{x}))^T \Big| \sum_{j=1}^m g_j(\mathbf{x}) = 0 \right\}$$

and assume $\mathbf{g}(\mathbf{x}) \in \mathcal{G}$. Zou et al. (2008) referred to the vectors in $\mathcal{G}$ as *multicategory margin vectors*.

Since a margin vector $\mathbf{g}(\mathbf{x})$ induces a classifier, we explore the minimization of ϵ with respect to (w.r.t.) $\mathbf{g}(\mathbf{x})$. However, this minimization problem is intractable because $\mathbb{I}_{[\phi(\mathbf{x})\neq c]}$ is the 0-1 function. Various tractable *surrogate loss* functions $\zeta(\mathbf{g}(\mathbf{x}), c)$ are thus used to approximate $\mathbb{I}_{[\phi(\mathbf{x})\neq c]}$. The corresponding population and empirical risk functions are given by

$$\mathcal{R}(P, \mathbf{g}) = E_X \Big(\sum_{c=1}^m \zeta(\mathbf{g}(\mathbf{x}), c) P_c(\mathbf{x}) \Big),$$

$$\hat{\mathcal{R}}(\mathbf{g}) = \frac{1}{n} \sum_{i=1}^n \zeta(\mathbf{g}(\mathbf{x}_i), c_i),$$

where $E_X(\cdot)$ is the expectation taken w.r.t. the distribution of X.

If α is a positive constant that does not depend on $(\mathbf{x}, c)$, $\text{argmin}_{\mathbf{g}(\mathbf{x})\in\mathcal{G}} \frac{1}{\alpha} \hat{\mathcal{R}}(\mathbf{g})$ is equivalent to $\text{argmin}_{\mathbf{g}(\mathbf{x})\in\mathcal{G}} \hat{\mathcal{R}}(\mathbf{g})$. We thus present the following definition.

Definition 1 *A surrogate loss $\zeta(\mathbf{g}(\mathbf{x}), c)$ is said to be a majorization of $\mathbb{I}_{[\phi(\mathbf{x})\neq c]}$ w.r.t. $(\mathbf{x}, c)$ if $\zeta(\mathbf{g}(\mathbf{x}), c) \geq \alpha \mathbb{I}_{[\phi(\mathbf{x})\neq c]}$ where α is a positive constant that does not depend on $(\mathbf{x}, c)$.*

Given a majorization function $\zeta(\mathbf{g}(\mathbf{x}), c)$, the classifier resulting from the minimization of $\hat{\mathcal{R}}(\mathbf{g})$ w.r.t. the margin vector $\mathbf{g}(\mathbf{x})$ is called a margin-based classifier or a margin-based classification method. Therefore, a margin-based classifier corresponds to a so-called majorization-minimization procedure. In the binary classification setting, a wide variety of classifiers can be understood as minimizers of a majorization loss function of the 0-1 loss. If such functions satisfy other technical conditions, the resulting classifiers can be shown to be Bayes consistent (Bartlett et al., 2006). It seems reasonable to pursue a similar development in the case of multicategory classification, and indeed such a proposal has been made by Zou et al. (2008) (see also Tewari and Bartlett (2007); Zhang (2004)). The following definition refines the definition of Zou et al. (2008). (Specifically, we do not require that the function $\zeta(\mathbf{g}(\mathbf{x}), c)$ depends only on $g_c(\mathbf{x})$.)

Definition 2 *A surrogate function $\zeta(\mathbf{g}(\mathbf{x}), c)$ is said to be* Fisher consistent w.r.t. *a margin vector $\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), \ldots, g_m(\mathbf{x}))^T$ at $\mathbf{x}$ if* (i) *the following risk minimization problem*

$$\hat{\mathbf{g}}(\mathbf{x}) = \underset{\mathbf{g}(\mathbf{x})\in\mathcal{G}}{\text{argmin}} \sum_{c=1}^m \zeta(\mathbf{g}(\mathbf{x}), c) P_c(\mathbf{x}) \tag{1}$$

has a unique solution $\hat{\mathbf{g}}(\mathbf{x}) = (\hat{g}_1(\mathbf{x}), \ldots, \hat{g}_m(\mathbf{x}))^T$; and (ii)

$$\underset{c}{\text{argmax}}\, \hat{g}_c(\mathbf{x}) = \underset{c}{\text{argmax}}\, P_c(\mathbf{x}).$$

2.2 Multicategory Losses

Zou et al. (2008) derived multicategory boosting algorithms by using $\zeta(\mathbf{g}(\mathbf{x}), c) = \exp(-g_c(\mathbf{x}))$. In their

discrete boosting algorithms, the margin vector $\mathbf{g}(\mathbf{x})$ is modeled as an m-vector function with one and only one positive element. In this case, $\mathbb{I}_{[g_c(\mathbf{x})\le 0]}$ is equal to $\mathbb{I}_{[\phi(\mathbf{x})\neq c]}$. Consequently, $\exp(-g_c(\mathbf{x}))$ is a majorization of $\mathbb{I}_{[\phi(\mathbf{x})\neq c]}$. Therefore, the discrete AdaBoost algorithms of Zou et al. (2008) still approximate the original empirical 0-1 loss function. However, in the general case, $\exp(-g_c(\mathbf{x}))$ is not a majorization of $\mathbb{I}_{[\phi(\mathbf{x})\neq c]}$. Thus the multicategory GentleBoost algorithm of Zou et al. (2008) is not a margin-based method.

Friedman et al. (2000) proposed a multicategory LogitBoost algorithm by using the negative multinomial log-likelihood function, which is given by

$$\begin{aligned}\mathcal{L}(\mathbf{g}(\mathbf{x}),c) &= \log\sum_{j=1}^{m}\exp\big(g_j(\mathbf{x})-g_c(\mathbf{x})\big) \\ &= \log\Big[1+\sum_{j\neq c}\exp(g_j(\mathbf{x})-g_c(\mathbf{x}))\Big]\end{aligned} \tag{2}$$

at $(\mathbf{x},c)$. Although $\log[1+\exp(-g_c(\mathbf{x}))]$ is an upper bound of $\log(2)\mathbb{I}_{[g_c(\mathbf{x})\le 0]}$, it is not a majorization of $\mathbb{I}_{[\phi(\mathbf{x})\neq c]}$. However, $\mathcal{L}(\mathbf{g}(\mathbf{x}),c)$ is a majorization of $\mathbb{I}_{[\phi(\mathbf{x})\neq c]}$ because of $\mathcal{L}(\mathbf{g}(\mathbf{x}),c)\ge\log(2)\mathbb{I}_{[\phi(\mathbf{x})\neq c]}$. Thus, the multicategory LogitBoost algorithm (Friedman et al., 2000) is also a margin-based method.

It is worth noting that $\log[1+\exp(-g_c(\mathbf{x}))]$ is the majorization of $\mathbb{I}_{[\phi(\mathbf{x})\neq c]}$ if the margin vector $\mathbf{g}(\mathbf{x})$ has only one positive element. Unfortunately, when this majorization as well as $\mathcal{L}(\mathbf{g}(\mathbf{x}),c)$ are used to derive multicategory discrete boosting algorithms, a closed-form solution no longer exists.

In the case of the multicategory SVM algorithm, Crammer and Singer (2001) used the surrogate:

$$\mathcal{H}(\mathbf{g}(\mathbf{x}),c)=\max\{g_j(\mathbf{x})+1-\mathbb{I}_{[j=c]}\}-g_c(\mathbf{x}). \tag{3}$$

It is easily seen that

$$\begin{aligned}\mathbb{I}_{[\phi(\mathbf{x})\neq c]} &= \mathbb{I}_{[\exists j\neq c,\ g_j(\mathbf{x})-g_c(\mathbf{x})>0]} \\ &\le \max\Big\{g_j(\mathbf{x})+1-\mathbb{I}_{[j=c]}\Big\}-g_c(\mathbf{x}).\end{aligned}$$

This shows that $\mathcal{H}(\mathbf{g}(\mathbf{x}),c)$ is a majorization of $\mathbb{I}_{[\phi(\mathbf{x})\neq c]}$, but it is Fisher consistent only when $\max_l P_l(\mathbf{x})>1/2$ (Zhang, 2004).

3 Coherence Functions

Since hinge-type loss functions are not smooth, existing multicategory SVMs do not directly estimate the class probability $P_c(\mathbf{x})$. Moreover, it is rare to devise a boosting algorithm with them. However, logistic regression extends naturally from binary classification to multicategory classification by simply using the multinomial likelihood in place of the binomial likelihood. In this section, we present a smooth and Fisher-consistent majorization loss, which bridges hinge-type losses and the negative multinomial log-likelihood. Thus, it is applicable to the construction of multicategory margin-based classifiers.

3.1 Definition

In order to obtain a majorization function of $\mathbb{I}_{[\phi(\mathbf{x})\neq c]}$, we express $\max\{g_j(\mathbf{x})+1-\mathbb{I}_{[j=c]}\}$ as $\sum_{j=1}^m\beta_j^c(\mathbf{x})\big[1+g_j(\mathbf{x})-\mathbb{I}_{[j=c]}\big]$ where

$$\beta_j^c(\mathbf{x})=\begin{cases}1 & j=\operatorname{argmax}_l\{g_l(\mathbf{x})+1-\mathbb{I}_{[l=c]}\}\\ 0 & \text{otherwise.}\end{cases}$$

Motivated by the idea behind deterministic annealing (Rose et al., 1990), we relax this hard function $\beta_j^c(\mathbf{x})$, retaining only $\beta_j^c(\mathbf{x})>0$ and $\sum_{j=1}^m\beta_j^c(\mathbf{x})=1$. With respect to a soft $\beta_j^c(\mathbf{x})$ respecting these constraints, we maximize $\sum_{j=1}^m\beta_j^c(\mathbf{x})\big[1+g_j(\mathbf{x})-\mathbb{I}_{[j=c]}\big]$ under an entropy constraint, namely,

$$\begin{aligned}\max_{\{\beta_j^c(\mathbf{x})\}}\Big\{F=&\sum_{j=1}^m\beta_j^c(\mathbf{x})\big[1+g_j(\mathbf{x})-\mathbb{I}_{[j=c]}\big]\\ &-T\sum_{j=1}^m\beta_j^c(\mathbf{x})\log\beta_j^c(\mathbf{x})\Big\},\end{aligned} \tag{4}$$

where we refer to $T>0$ as a temperature.

The maximization of F w.r.t. $\beta_j^c(\mathbf{x})$ is straightforward, and it gives rise to the following distribution

$$\beta_j^c(\mathbf{x})=\frac{\exp\big[\frac{1+g_j(\mathbf{x})-\mathbb{I}_{[j=c]}}{T}\big]}{\sum_l\exp\big[\frac{1+g_l(\mathbf{x})-\mathbb{I}_{[l=c]}}{T}\big]}. \tag{5}$$

The corresponding maximum of F is obtained by plugging (5) back into (4):

$$F^*=T\log\sum_j\exp\Big[\frac{1+g_j(\mathbf{x})-\mathbb{I}_{[j=c]}}{T}\Big].$$

Note that for $T>0$ we have

$$\begin{aligned}&T\log\Big[1+\sum_{j\neq c}\exp\frac{1+g_j(\mathbf{x})-g_c(\mathbf{x})}{T}\Big]\\ &=T\log\sum_j\exp(\frac{1+g_j(\mathbf{x})-\mathbb{I}_{[j=c]}}{T})-g_c(\mathbf{x})\\ &\ge\max_j\Big\{g_j(\mathbf{x})+1-\mathbb{I}_{[j=c]}\Big\}-g_c(\mathbf{x})\\ &\ge\mathbb{I}_{[\phi(\mathbf{x})\neq c]}.\end{aligned}$$

This thus leads us to the following family of majorization functions of $\mathbb{I}_{[\phi(\mathbf{x})\neq c]}$:

$$\mathcal{C}(\mathbf{g}(\mathbf{x}),c)=T\log\Big[1+\sum_{j\neq c}\exp\frac{1+g_j(\mathbf{x})-g_c(\mathbf{x})}{T}\Big],\ T>0. \tag{6}$$

We refer to the functions as *coherence functions* due to their statistical mechanical properties similar to those of deterministic annealing (Rose et al., 1990). Note that the coherence function is also a majorization of the multicategory hinge loss $\mathcal{H}(\mathbf{g}(\mathbf{x}), c)$ in (3).

When $T = 1$, we have

$$\mathcal{C}(\mathbf{g}(\mathbf{x}), c) = \log\Big[1+\sum_{j\neq c}\exp\big(1+g_j(\mathbf{x})-g_c(\mathbf{x})\big)\Big],$$

which is just an upper bound of the negative multinomial log-likelihood function $\mathcal{L}(\mathbf{g}(\mathbf{x}), c)$ in (2).

In the binary case, i.e. $m = 2$, we let $g_1(\mathbf{x}) = -g_2(\mathbf{x}) = \frac{1}{2}f(\mathbf{x})$ and encode $y = 1$ if $c = 1$ and $y = -1$ if $c = 2$. We can thus express the coherence function as

$$\mathcal{C}(yf(\mathbf{x})) = T\log\Big[1+\exp\frac{1-yf(\mathbf{x})}{T}\Big],\ T > 0. \quad (7)$$

Figure 1 depicts the coherence function ($T = 1$) and other common loss functions for $m = 2$.

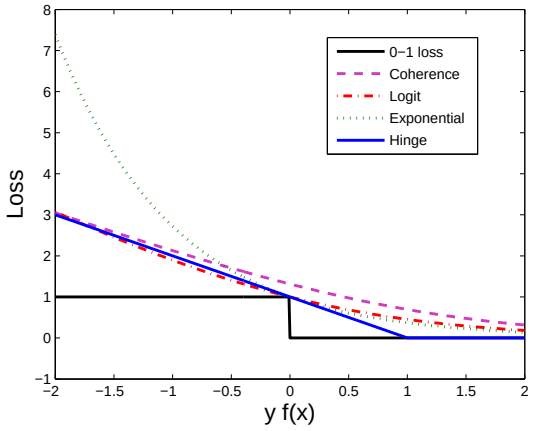


Figure 1: A variety of loss functions, which are regarded as a function of $yf(\mathbf{x})$. Here $T = 1$ in the coherence loss. *Logit loss*: $\frac{1}{\log 2}\log[1+\exp(-yf(\mathbf{x}))]$; *Exponential loss*: $\exp(-yf(\mathbf{x}))$; *Hinge loss*: $[1-yf(\mathbf{x})]_+$ where $[u]_+ = u$ if $u \geq 0$ and $[u]_+ = 0$ otherwise.

3.2 Properties

The following theorem shows that the coherence function is Fisher consistent.

Theorem 1 *Assume $P_c(\mathbf{x}) > 0$ for $c = 1, \ldots, m$. Consider the optimization problem*

$$\operatorname*{argmax}_{\mathbf{g}(\mathbf{x})\in\mathcal{G}}\sum_{c=1}^{m} T\log\Big[1+\sum_{j\neq c}\exp\frac{1+g_j(\mathbf{x})-g_c(\mathbf{x})}{T}\Big]P_c(\mathbf{x})$$

for a fixed $T > 0$ and let $\hat{\mathbf{g}}(\mathbf{x}) = (\hat{g}_1(\mathbf{x}), \ldots, \hat{g}_m(\mathbf{x}))^T$ be its solution. Then $\hat{\mathbf{g}}(\mathbf{x})$ is unique. Moreover, if $P_i(\mathbf{x}) < P_j(\mathbf{x})$, we have $\hat{g}_i(\mathbf{x}) < \hat{g}_j(\mathbf{x})$. Furthermore, after having obtained $\hat{\mathbf{g}}(\mathbf{x})$, $P_c(\mathbf{x})$ is given by

$$P_c(\mathbf{x}) = \frac{\sum_{l=1}^{m}\exp\frac{1+\hat{g}_l(\mathbf{x})+\hat{g}_c(\mathbf{x})-\mathbb{I}_{[l=c]}}{T}}{\sum_{j=1}^{m}\sum_{l=1}^{m}\exp\frac{1+\hat{g}_l(\mathbf{x})+\hat{g}_j(\mathbf{x})-\mathbb{I}_{[l=j]}}{T}}. \quad (8)$$

Moreover, we have the following properties.

Theorem 2 *Let $\mathcal{H}(\mathbf{g}(\mathbf{x}), c)$, $\beta_j^c(\mathbf{x})$ and $\mathcal{C}(\mathbf{g}(\mathbf{x}), c)$ be defined by (3), (5) and (6), respectively. Then,*

$$\mathcal{S}(\mathbf{g}(\mathbf{x}), c) \leq \mathcal{C}(\mathbf{g}(\mathbf{x}), c) - T\log m \leq \mathcal{H}(\mathbf{g}(\mathbf{x}), c),$$

where

$$\mathcal{S}(\mathbf{g}(\mathbf{x}), c) = \frac{1}{m}\sum_{j\neq c}\big(1+g_j(\mathbf{x})-g_c(\mathbf{x})\big).$$

Importantly, when treating $\mathbf{g}(\mathbf{x})$ fixed and considering $\beta_j^c(\mathbf{x})$ and $\mathcal{C}(\mathbf{g}(\mathbf{x}), c)$ as functions of T, we have

Theorem 3 *Under the conditions in Theorem 2, for a fixed $\mathbf{g}(\mathbf{x})$ we have*

(i) $\lim_{T\to\infty}\mathcal{C}(\mathbf{g}(\mathbf{x}), c) - T\log m = \mathcal{S}(\mathbf{g}(\mathbf{x}), c)$ *and*

$$\lim_{T\to\infty}\beta_j^c(\mathbf{x}) = \frac{1}{m} \quad \text{for } j = 1, \ldots, m$$

(ii) $\lim_{T\to 0}\mathcal{C}(\mathbf{g}(\mathbf{x}), c) = \mathcal{H}(\mathbf{g}(\mathbf{x}), c)$ *and*

$$\lim_{T\to 0}\beta_j^c(\mathbf{x}) = \begin{cases} 1 & j = \operatorname{argmax}_l\{g_l(\mathbf{x})+1-\mathbb{I}_{[l=c]}\} \\ 0 & \text{otherwise.}\end{cases}$$

It is worth noting that Theorem 3-(ii) shows that at $T = 0$, $\mathcal{C}(\mathbf{g}(\mathbf{x}), c)$ reduces to the multicategory hinge loss $\mathcal{H}(\mathbf{g}(\mathbf{x}), c)$, which is used by Crammer and Singer (2001).

As an immediate corollary of Theorems 2 and 3 in the binary case ($m = 2$), we have

Corollary 1 *Let $\mathcal{C}(yf(\mathbf{x}))$ be defined by (7). Then*

(i) $(1-yf(\mathbf{x}))_+ \leq \mathcal{C}(yf(\mathbf{x})) \leq T\log 2+[1-yf(\mathbf{x})]_+$;

(ii) $\lim_{T\to 0}\mathcal{C}(yf(\mathbf{x})) = [1-yf(\mathbf{x})]_+$;

(iii) $\frac{1}{2}(1-yf(\mathbf{x})) \leq \mathcal{C}(yf(\mathbf{x})) - T\log 2$;

(iv) $\lim_{T\to\infty}\mathcal{C}(yf(\mathbf{x})) - T\log 2 = \frac{1}{2}(1-yf(\mathbf{x}))$.

Graphs of $\mathcal{C}(yf(\mathbf{x}))$ with different values of T are shown in Figure 2. We can see that $\mathcal{C}(yf(\mathbf{x}))$ with $T = 0.01$ is almost the same as the hinge loss $[1-yf(\mathbf{x})]_+$.

Wang et al. (2005) derived an annealed discriminant analysis algorithm in which the loss function is

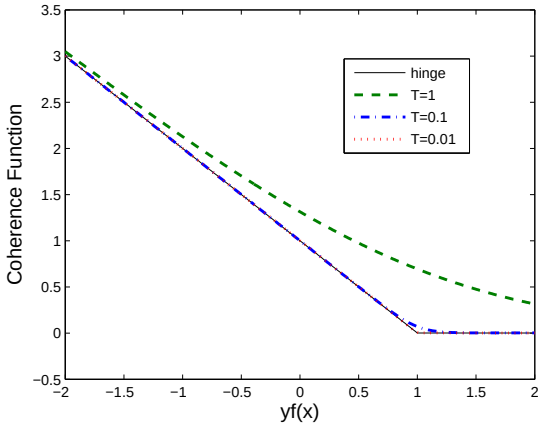


Figure 2: Coherence functions with $T = 1$, $T = 0.1$ and $T = 0.01$.

$$\mathcal{A}(\mathbf{g}(\mathbf{x}), c) = T \log \Big[1 + \sum_{j \neq c} \exp \frac{g_j(\mathbf{x}) - g_c(\mathbf{x})}{T}\Big],\ T > 0.$$

Thus, the negative multinomial log-likelihood function $\mathcal{L}(\mathbf{g}(\mathbf{x}), c)$ and the conventional logistic regression are respectively the special cases of $\mathcal{A}(\mathbf{g}(\mathbf{x}), c)$ and the annealed discriminant analysis algorithm with $T = 1$ (also refer to Zhang and Oles (2001) for the binary case). However, since

$$\lim_{T \longrightarrow 0} \mathcal{A}(\mathbf{g}(\mathbf{x}), c) = \max_j (g_j(\mathbf{x}) - g_c(\mathbf{x})),$$

it is no longer guaranteed that $\mathcal{A}(\mathbf{g}(\mathbf{x}), c)$ is always a majorization of $\mathbb{I}_{[\phi(\mathbf{x}) \neq c]}$ for any $T > 0$.

4 The GentleBoost Algorithm

In this section we apply the coherence function to the development of multicategory margin-based boosting algorithms. Like the negative multinomial log-likelihood function, when the coherence function is used to devise multicategory discrete boosting algorithms, a closed-form solution no longer exists. We instead use the coherence function to devise a genuine multicategory margin-based boosting algorithm. With a derivation similar to that in Friedman et al. (2000); Zou et al. (2008), our GentleBoost algorithm is shown in Algorithm 1.

5 Experimental Evaluation

We compare our algorithm (called `GentleBoost.C`) with some representative multicategory boosting algorithms, including `AdaBoost.MH` (Schapire and Singer, 1999), multicategory LogitBoost (`MulLogitBoost`) (Friedman et al., 2000) and multicategory GentleBoost (`GentleBoost.E`) (Zou et al., 2008), on six publicly available datasets (`Vowel`, `Waveform`, `Image Segmentation`, `Optdigits`, `Pendigits` and `Satimage`) from the UCI Machine Learning Repository. Following the settings in Friedman et al. (2000); Zou et al. (2008), we use predefined training samples and test samples for these six datasets. Summary information for the datasets is given in Table 1.

Based on the experimental strategy in Zou et al. (2008), eight-node regression trees are used as weak learners for all the boosting algorithms with the exception of `AdaBoost.MH`, which is based on eight-node classification trees. In the experiments, we observe that the performance of all the methods becomes stable after about 50 boosting steps. Hence, the number of boosting steps for all the methods is set to 100 ($H = 100$) in all the experiments. The test error rates (in %) of all the boosting algorithms are shown in Table 2, from which we can see that all the boosting methods achieve much better results than CART, and our method slightly outperforms the other boosting algorithms. Among all the datasets tested, `Vowel` and `Waveform` are the most difficult for classification. The notably better performance of our method for these two datasets reveals its promising properties. Figure 3 depicts the test error curves of `MulLogitBoost`, `GentleBoost.E` and `GentleBoost.C` on these two datasets.

Theorem 3 shows that as $T \to 0$, $\mathcal{C}(\mathbf{g}(\mathbf{x}), c)$ approaches $\max\{g_j(\mathbf{x}) + 1 - \mathbb{I}_{[j=c]}\} - g_c(\mathbf{x})$. This encourages us to try to decrease T gradually over the boosting steps. However, when T gets very small, it can lead to numerical problems and often makes the algorithm unstable. The experiments show that when T takes a value in $[0.1, 2]$, our algorithm is always able to obtain promising performance. Here our reported results are based on the setting of $T = 1$. Recall that $\mathcal{L}(\mathbf{g}(\mathbf{x}), c)$ is the special case of $\mathcal{A}(\mathbf{g}(\mathbf{x}), c)$ with $T = 1$, so the comparison of `GentleBoost.C` with `MulLogitBoost` is fair based on $T = 1$.

As we established in Section 2.2, `GentleBoost.E` does not implement a margin-based decision because the loss function used in this algorithm is not a majorization of the 0-1 loss. Our experiments show that `MulLogitBoost` and `GentleBoost.C` are competitive, and outperform `GentleBoost.E`.

6 Conclusion

In this paper, we have proposed a novel majorization function and a multicategory boosting algorithm based

Algorithm 1 GentleBoost.C($\{(\mathbf{x}_i, c_i)\}_{i=1}^n \subset \mathbb{R}^p \times \{1, \ldots, m\}$, T, H)

1: Start with uniform weights $w_{ij} = 1/n$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$, and $\beta_j(\mathbf{x}) = 1/m$ and $g_j(\mathbf{x}) = 0$ for $j = 1, \ldots, m$.

2: Repeat for $h = 1$ to H:

(a) Repeat for $j = 1, \ldots, m$:

(i) Compute working responses and weights in the jth class,

$$z_{ij} = \frac{\mathbb{I}_{[j=c_i]} - \beta_j(\mathbf{x}_i)}{\beta_j(\mathbf{x}_i)(1 - \beta_j(\mathbf{x}_i))},$$
$$w_{ij} = \beta_j(\mathbf{x}_i)(1 - \beta_j(\mathbf{x}_i)).$$

(ii) Fit the regression function $g_j^{(h)}(\mathbf{x})$ by a weighted least-squares fit of the working response z_{ij} to $\mathbf{x}_i$ with weights w_{ij} on the training data.

(iii) Set $g_j(\mathbf{x}) \leftarrow g_j(\mathbf{x}) + g_j^{(h)}(\mathbf{x})$.

(b) Set $g_j(\mathbf{x}) \leftarrow \frac{m-1}{m}\left[g_j(\mathbf{x}) - \frac{1}{m}\sum_{l=1}^m g_l(\mathbf{x})\right]$ for $j = 1, \ldots, m$.

(c) Compute $\beta_j(\mathbf{x}_i)$ for $j = 1, \ldots, m$ as

$$\beta_j(\mathbf{x}_i) = \begin{cases} \frac{\exp\left(\frac{1+g_j(\mathbf{x}_i)-g_{c_i}(\mathbf{x}_i)}{T}\right)}{1+\sum_{j\neq c_i}\exp\left(\frac{1+g_j(\mathbf{x}_i)-g_{c_i}(\mathbf{x}_i)}{T}\right)} & \text{if } j \neq c_i, \\ \frac{1}{1+\sum_{j\neq c_i}\exp\left(\frac{1+g_j(\mathbf{x}_i)-g_{c_i}(\mathbf{x}_i)}{T}\right)} & \text{if } j = c_i. \end{cases}$$

3: Output $\phi(\mathbf{x}) = \text{argmax}_j\, g_j(\mathbf{x})$.

Table 1: Summary of benchmark datasets.

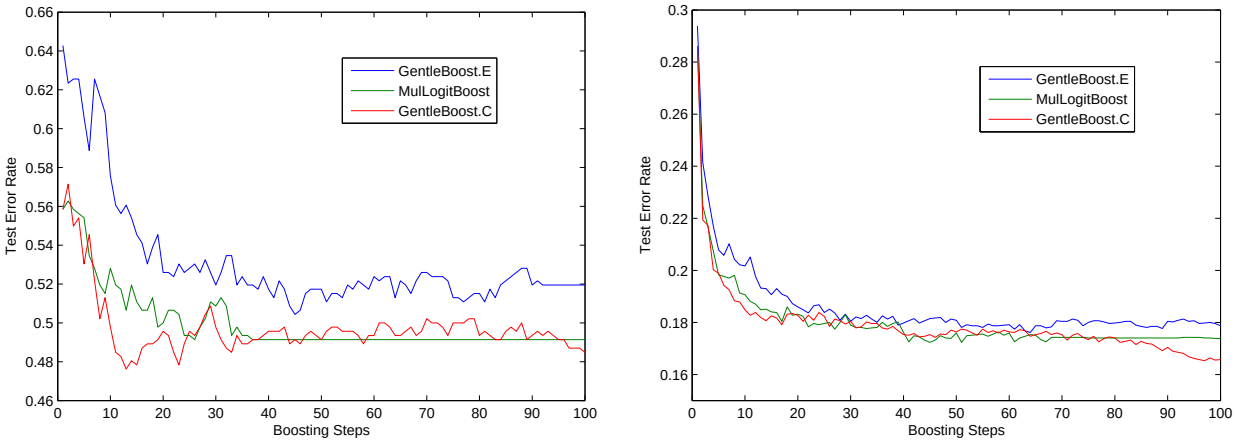
Dataset	# Train	# Test	# Features	# Classes
`Vowel`	528	462	10	11
`Waveform`	300	4700	21	3
`Segmentation`	210	2100	19	7
`Optdigits`	3823	1797	64	10
`Pendigits`	7494	3498	16	10
`Satimage`	4435	2000	36	6

Table 2: Test error rates of our method and related methods (in %). The best result for each dataset is shown in bold.

Dataset	`CART`	`AdaBoost.MH`	`MulLogitBoost`	`GentleBoost.E`	`GentleBoost.C`
`Vowel`	54.10	50.87	49.13	50.43	**47.62**
`Waveform`	31.60	18.22	17.23	17.62	**16.53**
`Segmentation`	9.80	5.29	4.10	4.52	**4.05**
`Optdigits`	16.60	5.18	3.28	5.12	**3.17**
`Pendigits`	8.32	5.86	**3.12**	3.95	3.14
`Satimage`	14.80	10.00	9.25	12.00	**8.75**

on this function. The majorization function is Fisher consistent, differential and convex. Thus, it is appropriate for the design of margin-based boosting algorithms. While our main focus has been theoretical, we have also shown experimentally that our boosting algorithm is effective, although it will be necessary to investigate its empirical performance more extensively.

Owing to the relationship of our majorization function with the hinge loss and the negative multinomial log-likelihood function, it is also natural to use the coherence function to devise a multicategory margin-based classifier as an alternative to existing multicategory SVMs and multinomial logistic regression models.



(a) Vowel (b) Waveform

Figure 3: Test error rates versus boosting steps.

A Proof of Theorem 1

A map $L : \Omega \to \mathbb{R}$, where Ω is a normal function space defined over $\mathbb{R}^d$, is said to be *Gateaux differentiable* at $g(\mathbf{x}) \in \Omega$, if for every fixed $h \in \Omega$ there exists

$$L'(g(\mathbf{x})) = \lim_{t \to 0} \frac{L(g(\mathbf{x}) + th) - L(g(\mathbf{x}))}{t}.$$

In our derivation, for notational simplicity, we omit $\mathbf{x}$ in the functions and denote $L'(g_j(\mathbf{x}))$ by $\frac{\partial L}{\partial g_j}$.

Without loss of generality, we let $T = 1$ in the following derivation. Consider the following Lagrangian

$$L = \sum_{j=1}^{m} \log \Big[1 + \sum_{l \neq j} \exp(1 + g_l - g_j)\Big] P_j + \lambda \sum_{j=1}^{m} g_j$$

and calculate the first and second derivatives of L w.r.t. the g_c as

$$\begin{aligned}
\frac{\partial L}{\partial g_c} &= -\frac{\sum_{l \neq c} \exp(1+g_l-g_c)}{1+\sum_{l \neq c} \exp(1+g_l-g_c)} P_c \\
&\quad + \sum_{j \neq c} \frac{\exp(1 + g_c - g_j)}{1 + \sum_{l \neq j} \exp(1+g_l-g_j)} P_j + \lambda \\
&= -\frac{\sum_{l=1}^{m} \exp(1+g_l-g_c)}{1+\sum_{l \neq c} \exp(1+g_l-g_c)} P_c \\
&\quad + \sum_{j=1}^{m} \frac{\exp(1 + g_c - g_j)}{1 + \sum_{l \neq j} \exp(1+g_l-g_j)} P_j + \lambda \\
&= -\sum_{l=1}^{m} \beta_{cl} P_c + \sum_{j=1}^{m} \beta_{jc} P_j + \lambda,
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 L}{\partial g_c \partial g_k} &= -\beta_{ck} P_c + \sum_{l \neq c} \beta_{cl} \beta_{ck} P_c - \beta_{kc} P_k \\
&\quad + \beta_{kc} \sum_{l \neq k} \beta_{kl} P_k - \sum_{j \neq c,k} \beta_{jc} \beta_{jk} P_j \\
&= -\sum_{j=1}^{m} \beta_{jc} \beta_{jk} P_j
\end{aligned}$$

for $k \neq c$, and

$$\frac{\partial^2 L}{\partial g_c \partial g_c} = \sum_{j=1}^{m} \beta_{jc}(1 - \beta_{jc}) P_j$$

where

$$\begin{aligned}
\beta_{cc} &= \frac{1}{1 + \sum_{l \neq c} \exp(1 + g_l - g_c)} \\
\beta_{cj} &= \frac{\exp(1 + g_j - g_c)}{1 + \sum_{l \neq c} \exp(1 + g_l - g_c)}.
\end{aligned}$$

We denote $\mathbf{\Delta}_j = \mathrm{diag}(\beta_{j1}, \beta_{j2}, \ldots, \beta_{jm})$ and $\boldsymbol{\beta}_j = (\beta_{j1}, \beta_{j2}, \ldots, \beta_{jm})^T$. The Hessian matrix is

$$\mathbf{H} \triangleq \frac{\partial^2 L}{\partial \mathbf{g}^T \partial \mathbf{g}} = \sum_{j=1}^{m} P_j (\mathbf{\Delta}_j - \boldsymbol{\beta}_j \boldsymbol{\beta}_j^T).$$

For any nonzero $\mathbf{u} \in \mathbb{R}^m$ subject to $\sum_{j=1}^{m} u_j = 0$, it is easily seen that

$$\mathbf{u}^T \mathbf{H} \mathbf{u} = \sum_{j=1}^{m} P_j \Big[\sum_{c=1}^{m} \beta_{jc} u_c^2 - (\sum_{c=1}^{m} \beta_{jc} u_c)^2 \Big] \geq 0.$$

Here we use the fact that u^2 is convex. Moreover, the above inequality is strictly satisfied for any nonzero

$\mathbf{u}$ with $\sum_{j=1}^m u_j = 0$. This shows that the optimization problem has a strictly local minimum point $\hat{\mathbf{g}}$. Again, we note that the Hessian matrix is positive semi-definite, so $\sum_{j=1}^m \log\Big[1+\sum_{l\neq j}\exp(1+g_l-g_j)\Big]P_j$ is convex. Thus, $\hat{\mathbf{g}}$ is also the global minimum point.

Now we prove that if $P_c > P_k$, then $\hat{g}_c > \hat{g}_k$. Since $\hat{\mathbf{g}}$ is the solution of equations $\frac{\partial L}{\partial g_c} = 0$, it immediately follows that $\lambda = 0$ by using $\sum_{c=1}^m \frac{\partial L}{\partial g_c} = 0$. Hence,

$$\frac{P_c\sum_{l=1}^m \exp(1+\hat{g}_l-\hat{g}_c)}{1+\sum_{l\neq c}\exp(1+\hat{g}_l-\hat{g}_c)} = \sum_{j=1}^m \frac{P_j\times\exp(1+\hat{g}_c-\hat{g}_j)}{1+\sum_{l\neq j}\exp(1+\hat{g}_l-\hat{g}_j)},$$

from which we get

$$\begin{aligned}\frac{P_c}{P_k} &= \frac{\exp(\hat{g}_c)}{\exp(\hat{g}_k)}\frac{\exp(\hat{g}_c)+\sum_{l\neq c}\exp(1+\hat{g}_l)}{\exp(\hat{g}_k)+\sum_{l\neq k}\exp(1+\hat{g}_l)} \qquad (9)\\ &= \frac{\exp(2\hat{g}_c)-\exp(1+2\hat{g}_c)+\exp(\hat{g}_c)\sum_{l=1}^m\exp(1+\hat{g}_l)}{\exp(2\hat{g}_k)-\exp(1+2\hat{g}_k)+\exp(\hat{g}_k)\sum_{l=1}^m\exp(1+\hat{g}_l)}\\ &> 1.\end{aligned}$$

Consequently,

$$\begin{aligned}0 &> \big[\exp(2\hat{g}_c)-\exp(2\hat{g}_k)\big]\big[1-\exp(1)\big]\\ &\quad+\big[\exp(\hat{g}_c)-\exp(\hat{g}_k)\big]\sum_{l=1}^m\exp(1+\hat{g}_l)\\ &= (\exp(\hat{g}_c)-\exp(\hat{g}_k))\Big[\exp(\hat{g}_c)+\exp(\hat{g}_k)\\ &\quad+\sum_{l\neq c,k}\exp(1+\hat{g}_l)\Big].\end{aligned}$$

Thus we obtain $\hat{g}_c > \hat{g}_k$. From (9), we get (8).

B Proof of Theorem 2

First, consider that

$$\begin{aligned}&T\log m+\mathcal{S}(\mathbf{g}(\mathbf{x}),c)-\mathcal{C}(\mathbf{g}(\mathbf{x}),c)\\ &= T\log\frac{m\exp\frac{1}{m}\sum_{j\neq c}\frac{1+g_j(\mathbf{x})-g_c(\mathbf{x})}{T}}{1+\sum_{j\neq c}\exp\frac{1+g_j(\mathbf{x})-g_c(\mathbf{x})}{T}}\\ &\leq T\log\frac{1+\sum_{j\neq c}\exp\frac{1+g_j(\mathbf{x})-g_c(\mathbf{x})}{T}}{1+\sum_{j\neq c}\exp\frac{1+g_j(\mathbf{x})-g_c(\mathbf{x})}{T}} = 0.\end{aligned}$$

Here we use the fact that $\exp(\cdot)$ is convex. Second, assume that $l = \mathrm{argmax}_j\{g_j(\mathbf{x})+1-\mathbb{I}_{[j=c]}\}$. Then

$$\begin{aligned}&T\log m+\mathcal{H}(\mathbf{g}(\mathbf{x}),c)-\mathcal{C}(\mathbf{g}(\mathbf{x}),c)\\ &= T\log\frac{m\exp\frac{1+g_l(\mathbf{x})-g_c(\mathbf{x})-\mathbb{I}_{[l=c]}}{T}}{1+\sum_{j\neq c}\exp\frac{1+g_j(\mathbf{x})-g_c(\mathbf{x})}{T}} \geq 0.\end{aligned}$$

Acknowledgements

Zhihua Zhang is supported in part by program for Changjiang Scholars and Innovative Research Team in University (IRT0652, PCSIRT), China. Zhang, Li and Yeung are supported by General Research Fund 621407 from the Research Grants Council of the Hong Kong Special Administrative Region, China. Michael Jordan acknowledges support from NSF Grant 0509559 and grants from Google and Microsoft Research.

References

Allwein, E. L., R. E. Schapire, and Y. Singer (2000). Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research 1*, 113–141.

Bartlett, P. L., M. I. Jordan, and J. D. McAuliffe (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association 101*(473), 138–156.

Cortes, C. and V. Vapnik (1995). Support-vector networks. *Machine Learning 20*, 273–297.

Crammer, K. and Y. Singer (2001). On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research 2*, 265–292.

Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Information and Computation 21*, 256–285.

Freund, Y. and R. E. Schapire (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences 55*(1), 119–139.

Friedman, J. H., T. Hastie, and R. Tibshirani (2000). Additive logistic regression: A statistical view of boosting. *The Annals of Statistics 28*(2), 337–374.

Rose, K., E. Gurewitz, and G. C. Fox (1990). Statistical mechanics and phase transitions in clustering. *Physics Review Letters 65*, 945–948.

Schapire, R. E. and Y. Singer (1999). Improved boosting algorithms using confidence-rated predictions. *Machine Learning 37*, 297–336.

Tewari, A. and P. L. Bartlett (2007). On the consistency of multiclass classification methods. *Journal of Machine Learning Research 8*, 1007–1025.

Wang, G., Z. Zhang, and F. H. Lochovsky (2005). Annealed discriminant analysis. In *ECML*.

Zhang, T. (2004). Statistical analysis of some multicategory large margin classification methods. *Journal of Machine Learning Research 5*, 1225–1251.

Zhang, T. and F. Oles (2001). Text categorization based on regularized linear classification methods. *Information Retrieval 4*, 5–31.

Zou, H., J. Zhu, and T. Hastie (2008). New multicategory boosting algorithms based on multicategory Fisher-consistent losses. *Annals of Applied Statistics 2*(4), 1290–1306.