2009－CVIM－166 (11)
2009／3／13

アンサンブル型最小分類誤り学習の提案

渡辺 秀行[†] 片桐 滋[††] 山田 幸太[††] 中村 篤[†††]
マクダーモット エリック[†††] 渡部 晋治[†††] 谷口 真一[††] 西島 奈甫[††††]
大崎 美穂[††††]

† 情報通信研究機構 MASTER プロジェクト 音声コミュニケーショングループ/ ATR 音声言語コミュニケーション研究所 〒619-0288 京都府相楽郡精華町光台 2-2-2
†† 同志社大学大学院 工学研究科 情報工学専攻
†††† 同志社大学理工学部 〒610-0394 京都府京田辺市多々羅都谷 1-3
††† 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
〒619-0237 京都府相楽郡精華町光台 2-4
E-mail: †hwatana@atr.jp or hideyuki.watanabe@nict.go.jp

あらまし 統計的パターン認識における究極の目標であるベイズ誤り推定との一貫性を持ちながら複数の弱分類器を統合する，アンサンブル型の最小分類誤り（MCE）学習法を提案する．まず、アンサンブル型学習法として注目を浴びているブースティングの，損失最小化としての最適性およびベイズ誤り推定との関連を述べる．そして MCE 学習の基本概念を概説し，損失最小化の解析を通して，ブースティングと MCE との密接な関連性を明らかにする．その上で，一般的な多クラス分類におけるベイズ誤り推定と一貫したアンサンブル型学習法である，アンサンブル型最小分類誤り（Ensemble-based MCE）学習法を提案する．

キーワード アンサンブル，最小分類誤り，MCE, ブースティング，ベイズ誤り

A Proposal of Ensemble-based Minimum Classification Error Training

Hideyuki WATANABE[†], Shigeru KATAGIRI[††], Kohta YAMADA[††], Atsushi NAKAMURA[†††], Erik MCDERMOTT[†††], Shinji WATANABE[†††], Shin'ichi TANIGUCHI[††], Naho NISHIJIMA[††††], and Miho OHSAKI[††††]

† MASTER Project, National Institute of Information and Communications Technology/ ATR Spoken Language Communication Research Labs. 2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288 Japan
†† Graduate School of Engineering, Doshisha University
†††† Faculty of Science and Engineering, Doshisha University
1-3 Tatara Miyakodani, Kyotanabe City, Kyoto 610-0394 Japan
††† NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237 Japan
E-mail: †hwatana@atr.jp or hideyuki.watanabe@nict.go.jp

Abstract We propose an ensemble-based minimum classification error (MCE) training method to combine multiple weak classifiers in a manner consistent with the ultimate standard, Bayes error estimation. First, we discuss boosting, a key methodology of ensemble training, from the viewpoints of mathematical optimality for loss minimization and its relationship to the Bayes error estimation. We also review the basic concept of MCE training, and elucidate the relationship between boosting and MCE by analyzing their loss minimization procedures. We then propose an ensemble-based training method named Ensemble-based MCE, which in principle leads to the Bayes error condition for a general multi-class task.

Key words ensemble, minimum classification error, MCE, boosting, Bayes error

1. はじめに

統計的パターン認識において目指すべき分類器の設計目標は，無限個の入力パターンに対する最小分類誤り確率状態の達成，すなわちベイズ誤りの推定である[1]～[3]．しかし現実的には，分類器の学習に用いられる学習標本は有限個である．音声認識などにおける複雑な認識課題において，学習標本数に対する学習パラメータ数が大きくなりすぎると，学習標本に対する正解率が高くても未知標本に対する正解率が低くなる．この問題に対する一般の解決策は学習パラメータ数を減らすことであるが，やみくもにパラメータ数を減らすだけでは，正解率そのものを大きく劣化させてしまう．

そこで，パラメータ数の少ない単純な分類器を多数学習し，それらの出力結果を統合して分類判断を行う，アンサンブル型の分類方式が提案され，その基本的な有効性が示されている[4]～[16]．中でも，個々の単純な分類器の不十分さを互いに補うように学習を行うブースティング（*boosting*）が大きな注目を浴びており，その軽量性と分類精度向上における効果が実証されつつある[8]～[16]．ブースティングでは，単純な分類器を弱分類器(注1)，そのアンサンブルを統合分類器(注2)といい，基本的に2値判別（2クラス分類）を扱う．標準的なブースティングとして定着しているのは，適応学習型の *AdaBoost*[9] である．Friedmanらは，AdaBoostがマージン(注3)の指数関数で定義される損失（指数損失）の平均の最小化を目指すことを示した[11]．更にMasonらは，マージンの微分可能な任意の単調減少損失関数の平均を，関数空間上の勾配降下法により最小化する，一般的なブースティングの枠組みである *MarginBoost* を定型化し，AdaBoostがその特殊例であることを示した[12]．ところがブースティングでは一般に，最小化の標的となる損失（指数損失など）と分類誤り数との関連づけが不十分であり，その結果として統合分類器の最適状態とベイズ誤りとの関係もまた不明瞭であった．

一方，分類器の学習理論として，最小分類誤り（*MCE*）学習法が提案されており，高い分類率を実現する学習法として広く浸透している[18]～[23]．この技術は，主に音声パターン認識の研究分野で発展してきたものであるが，本来は広範な認識分野においても利用可能な汎用的概念である．MCE学習は（2値判別より一般的な）多クラス分類におけるベイズ誤り推定を直接的に目指す．MCE学習では，各クラスの帰属度を測る判別関数を定め，学習標本に対する分類の誤り度合を表現する測度である誤分類測度（misclassification measure）と，この測度に対する平滑化された分類誤り数であるMCE損失を定式化し，MCE損失の平均を最小にするように分類器パラメータを学習する．MCE損失の利用は，その平滑化度合を制御することによって損失と分類誤り確率とを直接的に結びつけ，その結果として，分類誤り確率最小化という学習の本来目標と一貫性を持つ認識器設計を可能とする．MCE学習においても，未知標本に対する精度向上の課題が残されている．対処法として，MCE損失の平滑化度合の制御[18],[21]や，分類に対して本質的なクラス特徴のMCEによる探索[22],[23]などが検討されているが，未だに十分な解決には至っていない．

以上の考察より，ブースティングが持つ軽量性及び平滑効果とMCE学習が持つベイズ誤り推定との一貫性をうまく融合することにより，未知標本に対する分類精度が高い認識器が実現可能と期待される．そこで本稿では，異なる分野で発展した2つのパターン認識方法論，すなわちブースティングとMCEとの関連性を明らかにし，MCEに基づくアンサンブル型の学習法を提案する．まず2値判別において，MCE学習における誤分類測度とマージンの負が一致することから，MarginBoostにおける損失としてMCE損失をとることにより，MarginBoostがMCE学習の範疇に入ることを示す．続いて多クラス分類において，多クラス弱分類器を逐次的に増加させながら，多クラス判別関数のアンサンブルに対して直接的にMCE学習を適用する，アンサンブル型最小分類誤り（*Ensemble-based MCE*）学習法を提案する．これにより，多クラス分類におけるベイズ誤り推定と一貫したアンサンブル分類器の学習を定型化することができるとともに，弱分類器のアンサンブルがもつ平滑効果を活かせるものと期待される．

2. 統計的パターン認識の基礎

2.1 多クラス分類問題とベイズ誤り

$x \in \mathcal{X}$ を入力パターン(注4)，$y \in \mathcal{Y}$ を出力クラスラベルとする．なお $\mathcal{X}$ は入力パターン集合，また $\mathcal{Y} = \{1, 2, ..., U\}$ であり，U はクラス数である．パターン分類器は，$\mathcal{C} : \mathcal{X} \to \mathcal{Y}$ で表わされる写像として定義され，一般に次式の形式を有する．

$$\mathcal{C}(x) = u \ \text{ iff } \ g_u(x, \Lambda) = \max_{j \in \mathcal{Y}} g_j(x, \Lambda) \tag{1}$$

ここで $g_i(\cdot, \Lambda) : \mathcal{X} \to \mathbf{R}$（$\mathbf{R}$ は全実数集合）は i 番目クラスにおける判別関数であり，$g_i(x, \Lambda)$ は x の i 番目クラスへの帰属度を表す．Λ はパターン分類器の学習パラメータ集合である．明らかに，パターン分類器の性能を左右するのは判別関数 $g_1, ..., g_U$ である．

すべての入力に対する $\mathcal{C}$ の分類誤り確率は次式で与えられる．

$$\mathcal{E} = \mathrm{E}_{\mathcal{X},\mathcal{Y}}\left(1[g_y(x, \Lambda) \neq \max_{j \in \mathcal{Y}} g_j(x, \Lambda)]\right) \tag{2}$$

ここで $1[p]$ は命題 p が真なら1，偽なら0を返す関数であり，$\mathrm{E}_{\mathcal{X},\mathcal{Y}}$ は確率変数の組 x, y に関する期待値を表す．$\mathcal{E}$ を $g_1, ..., g_U$ の汎関数と捉えたとき，その最小値がベイズ誤りとなる．明らかに，すべてのパターンに対するベイズ誤り状態（最小分類誤り確率状態）をもたらす $g_1, ..., g_U$ の達成が望まれる．しかし現実的には，有限個（N 個）の教師付き学習標本集合 $\Omega_N = \{(x_n, y_n)\}_{n=1}^{N}$ のみを用いて，有限個のパラメータ

(注1)：弱識別器，弱仮説などともいう．

(注2)：統合仮説，強分類器，コミッティなどともいう．

(注3)：本稿ではマージンを，類境界とパターン標本との距離を表す幾何マージンではなく，判別関数値の大きさ等から成る関数マージンとする[17]．

(注4)：x は通常，分類のためのパターン特徴を表現するために観測値を変換した特徴量であり，固定次元パターンでも，音声特徴ベクトル系列のような可変次元パターンでもよい．

Λ が推定される．ここで $x_n \in \mathcal{X}$ は n 番目の学習パターンで，$y_n \in \mathcal{Y}$ は x_n に対する教師ラベルである．このときの学習目標は，式 (2) を有限の Ω_N で近似した次式の経験的誤り率である．

$$\mathcal{E}_0 = \frac{1}{N}\sum_{n=1}^{N} 1[g_{y_n}(x_n,\Lambda) \neq \max_{j\in\mathcal{Y}} g_j(x_n,\Lambda)] \tag{3}$$

システムを複雑にする（Λ に含まれるパラメータ数を多くする）ことにより，有限の Ω_N に対する $\mathcal{E}_0$ を任意に小さくすることができるが，$\mathcal{E}_0$ の最小状態は，本来目標とするべき，未知標本も含んだ $\mathcal{E}$ の最小状態すなわちベイズ誤りを意味しない．それればかりか，有限標本に対して $\mathcal{E}_0$ の最小化を追求しすぎると，かえって $\mathcal{E}$ が増大してしまう．これが過学習問題である．

2.2 2値判別問題とマージン

基本的なブースティングではクラスラベルが 1 と -1 である 2 値判別を考える．すなわち $\mathcal{Y} = \{1,-1\}$ と定義され，判別関数が $g_1(x,\Lambda)$，$g_{-1}(x,\Lambda)$ となる．更に 2 クラス共通の単一判別関数 $f(\cdot,\Lambda) : \mathcal{X} \to \mathbf{R}$ を $f(x,\Lambda) = g_1(x,\Lambda) - g_{-1}(x,\Lambda)$ で定義すると，分類規則は以下で与えられる．

$$C(x) = \mathrm{sgn}\big(f(x,\Lambda)\big) \tag{4}$$

ここで sgn は引数が 0 以上ならば 1，0 未満ならば -1 を返す関数である．このときの分類誤り確率 $\mathcal{E}$ および経験的分類誤り率 $\mathcal{E}_0$ は，式 (2)，(3) より次式となる．

$$\mathcal{E} = \mathrm{E}_{\mathcal{X},\mathcal{Y}}\Big(1[y \neq \mathrm{sgn}\big(f(x,\Lambda)\big)]\Big) \tag{5}$$

$$\mathcal{E}_0 = \frac{1}{N}\sum_{n=1}^{N} 1[y_n \neq \mathrm{sgn}\big(f(x_n,\Lambda)\big)] \tag{6}$$

2 値判別において，$f(x) = 0$ となる x の集合は f による類決定境界を表し，$f(x)$ が大きな正値（絶対値の大きな負値）をとるほど，x はよりクラス 1（-1）らしいと判断される．x の属する正しいクラスを y^* ($\in \{1,-1\}$) とすると，$z = y^* f(x)$ をマージンとよび，$z > 0$ は正しい分類を，$z < 0$ は誤分類を表す．また z は，大きいほど x の分類がより正確であり，0 に近いほど x が類決定境界付近に存在し，更に負の方向に大きいほど x が “かなり” 間違って分類されていることを意味する．

3. 基本的ブースティングアルゴリズム

3.1 アンサンブル型手法とブースティング

2.1 で述べたような過学習の解決策として，アンサンブル型の分類手法が開発されている [4]〜[16]．この手法は，複数の（多数の）単純な分類器を学習し，それらの分類結果を総合して最終判断を行う．単純な分類器の下す判断が互いに相関が小さければ，これらのアンサンブルをとることにより平滑効果による高い頑健性が得られることが知られている [4]〜[7]．中でも我々は，理論的整備が進んでおり注目を集めているブースティングに焦点を当てる [8]〜[16]．

ブースティングは基本的に 2 値判別問題を扱い，判別関数は以下のアンサンブル型の構造で与えられる．

$$F(x) = a_1 f_1(x) + a_2 f_2(x) + \cdots + a_T f_T(x) \tag{7}$$

1. Let $t = 0$, and initialize the weights $w_n = 1/N$ ($n = 1,2,...,N$) and $F_0 = 0$.
2. For $t = 1,2,...,T$ do:
   (a) Train a weak classifier $f = f_t \in \mathcal{F}$ for minimizing the $\boldsymbol{w}$-weighted empirical error rate over the training set Ω_N: $L_{\boldsymbol{w}}(f) = \sum_{n=1}^{N} w_n 1[y_n \neq f(x_n)]$.
   (b) Compute $\mathrm{err}_t = \sum_{n=1}^{N} w_n 1[y_n \neq f_t(x_n)]$.
   (c) Compute $a_t = (1/2)\log\big((1-\mathrm{err}_t)/\mathrm{err}_t\big)$.
   (d) Update the weights ($n = 1,2,...,N$):
   $$w_n \leftarrow w_n \cdot \exp\left(2a_t 1[y_n \neq f_t(x_n)]\right)/Z.$$
   (e) Update $F_t = F_{t-1} + a_t f_t$.
3. Return F_T ($= a_1 f_1 + a_2 f_2 + \cdots + a_T f_T$).

図 1　AdaBoost

ここで各 $f_t : \mathcal{X} \to \{1,-1\}$ ($t = 1,...,T$) (注5) は弱分類器であり，単純な構造の判別関数である．f_t は学習パラメータ数の少ない関数の集合 $\mathcal{F}$ から選ばれ，分類精度はあまり高くない．$\mathcal{F}$ は基本仮説集合とよばれる．a_t は f_t が下す判断の重要度を表す実数であり，分類器重みとよばれ，f_t の分類精度が高いほどより大きな値に設定される．F は統合分類器であり，最終分類判断は $x \mapsto \mathrm{sgn}\big(F(x)\big)$ で行われる．学習は基本的に，t に関して $f_t \in \mathcal{F}$ と a_t が逐次的に最適化され（t をステージとよぶ），統合分類器が $F_t = a_1 f_1 + \cdots + a_t f_t$ ($t = 1,2,...$) と増加的に構成されていく．弱分類器 f_t は，以前のステージで得られた弱分類器 $f_1,...,f_{t-1}$ の分類が困難であったパターン群に対して補強するように学習される．これにより，弱分類器が下す判断の不十分さを補いながら統合分類器が相補的かつ高精度な総合判断を下すことができるとともに，弱分類器同士の学習の相関が小さく，先述のようにアンサンブルの平滑効果が得られる．

3.2 AdaBoost

標準的なブースティングである AdaBoost [9] の手順を図 1 に示す．実数 Z は，$\sum_{n=1}^{N} w_n = 1$，$0 \leqq w_n \leqq 1$ ($\forall n$) をみたすための正規化パラメータである．$\boldsymbol{w} = \{w_n\}_{n=1}^{N}$ は観測重みとよばれ，各 x_n に対する学習の重点度を表すパラメータであり，w_n は Ω_N を標本空間とした x_n の相対度数である．2(a) における目的関数 $L_{\boldsymbol{w}}(f)$ は Ω_N に対する重み付き誤り率とよばれる．2(b)(c) において f_t に対する重み a_t が計算されるが，重み付き誤り率 err_t が小さいほど a_t がより大きな値に設定されるのがわかる．そして 2(d) において，f_t により誤分類された標本に対して観測重みがより大きな値に更新され，次ステージに対して，弱分類器が判断誤りを起こした標本群に対する重点度が増す機構となっている．

3.3 AdaBoost の最適性と MarginBoost

Friedman らは，損失最小化の概念を導入し，“指数損失” の最小化という観点から AdaBoost の最適性を明らかにした [11]．彼らは，AdaBoost における学習手順が，統合分類器 F が式 (7) に示される加法的モデルに従うという条件の下で，F のマージンの単調減少関数である指数損失関数 e^{-yF} の平均値

(注5)：f_t を連続的な実数値とするブースティングも存在する [10], [11].

1. Let $t=0$, and initialize $F_0=0$.
2. For $t=1,2,...,T$ do:
   (a) Compute the weights: $w_n = c'\big(y_n F_{t-1}(x_n)\big)\big/Z$ $(n=1,2,...,N)$.
   (b) Train a weak classifier $f=f_t\in\mathcal{F}$ for minimizing the $\boldsymbol{w}$-weighted empirical error rate over the training set Ω_N: $L_{\boldsymbol{w}}(f)=\sum_{n=1}^{N} w_n 1[y_n \neq f(x_n)]$.
   (c) Find the optimal classifier weight by line search: $a_t=\arg\min_a C(F_{t-1}+af_t)$.
   (d) Update $F_t=F_{t-1}+a_t f_t$.
3. Return F_T $(=a_1f_1+a_2f_2+\cdots+a_Tf_T)$.

図 2　MarginBoost

$\frac{1}{N}\sum_{n=1}^{N} e^{-y_nF(x_n)}$ を最小にする f_t, a_t $(t=1,...,T)$ を，前向きステージワイズで探索するものであることを示した．ここで前向きステージワイズの探索とは，t 番目ステージにおいて，学習対象である F を $F_t=a_1f_1+\cdots+a_tf_t$ とし，f_τ, a_τ $(\tau=1,...,t-1)$ を固定して f_t, a_t のみに関して最適化を行い，これを t に関して逐次的かつ増加的に実行することである．

しかし彼らの証明には，逐次的な f_t, a_t の最適化が全体的な F の最適性を保証するのかという疑問と，目的関数としての指数損失の妥当性の疑問がある．Mason らは，関数空間における勾配降下探索による損失最小化の視点から，この疑問を解消する道を与えた [12]．関数空間として，$\mathcal{F}$ の要素の線形結合となる関数の集合：$\mathrm{lin}\mathcal{F}$ を考え，その上の内積を $\langle h_1,h_2\rangle=\frac{1}{N}\sum_{n=1}^{N} h_1(x_n)h_2(x_n)$ $(h_1,h_2\in\mathrm{lin}\mathcal{F})$ で定める(注6)．ただし $\{(x_n,y_n)\}_{n=1}^{N}=\Omega_N$．ここで統合分類器 F が必ず $\mathrm{lin}\mathcal{F}$ に属することに注意．$c(z)$ をマージン z の単調減少かつ微分可能な任意の損失関数とし，F の評価汎関数 $C(F)$ を

$$C(F)=\frac{1}{N}\sum_{n=1}^{N} c\big(y_nF(x_n)\big)\quad (F\in\mathrm{lin}\mathcal{F}) \tag{8}$$

で定める．そして関数空間 $\mathrm{lin}\mathcal{F}$ において，$C(F)$ を $F\in\mathrm{lin}\mathcal{F}$ に関して最急降下法に基づき最小化する．これが MarginBoost であり，具体的手続きを図 2 に示す．

図 2 の 2(a) において，観測重み w_n は F_{t-1} に関して x_n が被る損失の "勾配" で与えられる．この $\boldsymbol{w}=\{w_n\}_{n=1}^{N}$ の下で，2(b) における最適な f_t は，$\mathcal{F}$ の要素の中で，F_{t-1} における $C(F)$ の最急降下方向 $-\nabla C(F_{t-1})$ と "できるだけ同じ方向を向いている" 関数である(注7)（詳細は [12] を参照）．そして 2(c) は "降下方向" f_t の上での直線探索であり，容易に示されるとおり，多くの学習標本が f_t により正しく分類されるならば，a_t は大きな値をとる．

3.4　分類誤り数損失の最小化としての MarginBoost

MarginBoost における損失 c として具体的なものを定めることにより，各種のブースティングが導出される．特に c を指数関数 e^{-z} に定めた場合，MarginBoost は AdaBoost に帰着する．このとき，負のマージンに対する損失の勾配の絶対値が指数関数的に増大し，F_{t-1} により大きく誤分類された "外れ値" に対して非常に大きな重みがかかることとなり，過学習を引き起こす．そこでこの現象を抑えた損失関数が検討されているが [11], [13]〜[15]，それらは（指数損失も含めて）いずれも分類誤り数を表すものではなく，分類誤り確率最小化としてのブースティングの最適性に欠ける．

ここで損失関数 c を $c_{\mathrm{ec}}(z)=1[z<0]$（分類誤り数損失）とすれば，式 (8) の $C(F)$ が式 (6) の $\mathcal{E}_0$ $(f=F$ とする$)$ と一致し，MarginBoost を分類誤り確率の最小化と直結させることができる．だが c_{ec} では，その不連続性ゆえに $C(F)$ の勾配を求めることができず，微分可能な関数である

$$c_{\mathrm{sig}}(z)=\frac{1}{1+\exp\big(\zeta_1 z\big)}\qquad (\zeta_1>0) \tag{9}$$

(sigmoid 関数) によって c_{ec} を近似的に置き換えることが考えられる．実際 Sano らは，c_{ec} を微分可能な関数で近似する方法を考案している [14], [15] が，c_{ec}（とその近似）の外れ値に対するロバスト性の改善に主眼がおかれ，ベイズ誤り推定との関連性が十分に検証されていない．

4.　最小分類誤り（MCE）学習との関連

式 (9) の c_{sig} は，MCE 学習 [18]〜[23] で定義される損失と本質的に同じものである．この事実から，MCE 学習と MarginBoost との密接な関係を明らかにすることができ，ベイズ誤り推定を目指したアンサンブル分類器の定型化への道が開ける．

4.1　MCE 学習の基礎

MCE 学習は，多クラス分類におけるベイズ誤り推定を直接的に目指した，判別関数学習法である．以下にその概要を示す．

MCE 学習の設計目的は基本的に 3 段階で構成される．まず 2.1 で述べたように，各クラスごとに判別関数 $\{g_i(x,\Lambda)\}_{i=1}^{U}$ を定める．分類規則は式 (1) で与えられる．次に，$\{g_i(x,\Lambda)\}_{i=1}^{U}$ の分類誤り度合を表す誤分類測度 $d_k(x,\Lambda)$ を定義する．ここで k は x が属する正しいクラスラベルである．種々の可能性の中から，たとえば次式により定義される．

$$d_k(x,\Lambda)=-g_k(x,\Lambda)+\log\left[\frac{1}{U-1}\sum_{j,j\neq k}\exp\Big\{\eta g_j(x,\Lambda)\Big\}\right]^{1/\eta}\qquad (\eta>0) \tag{10}$$

十分大きな η に対して，$d_k(x,\Lambda)<0$ は正しい分類に，$d_k(x,\Lambda)>0$ は誤分類に対応する(注8)．$d_k(x,\Lambda)$ が 0 に近いほど，x が類境界付近に存在することを示す．最後に $d_k(x,\Lambda)$ を基に，x に対する平滑化された分類誤り数損失（MCE 損失）$\ell\big(d_k(x,\Lambda)\big)$ を定義する．ここでも種々の可能性があるが，次式のように sigmoid 関数を使って定式化できる．

$$\ell\big(d_k(x,\Lambda)\big)=\frac{1}{1+\exp\big(-\zeta d_k(x,\Lambda)\big)}\qquad (\zeta>0) \tag{11}$$

(注6)：$\mathrm{lin}\mathcal{F}$ の各要素（関数）は抽象ベクトルである．

(注7)：一般に $-\nabla C(F_{t-1})\notin\mathcal{F}$ であることに注意．

(注8)：$\eta\to\infty$ とすれば $d_k(x,\Lambda)=-g_k(x,\Lambda)+\max_{j,j\neq k} g_j(x,\Lambda)$ となり，この対応関係が明快になる．

$\ell\big(d_k(x,\Lambda)\big)$ は $d_k(x,\Lambda)$ の単調増加関数であり，$\zeta\to\infty$ において，$d_k(x,\Lambda)<0$（正分類）ならば $\ell\to 0$，$d_k(x,\Lambda)>0$（誤分類）ならば $\ell\to 1$ となる．したがって MCE 損失は分類誤り数損失と直結しており，しかも Λ に関して微分可能である．最終的な設計目的は，次式の経験的平均損失の Λ による最小化である（ただし $\{(x_n,y_n)\}_{n=1}^N=\Omega_N$）．

$$L(\Lambda)=\frac{1}{N}\sum_{n=1}^{N}\ell\big(d_{y_n}(x_n,\Lambda)\big) \tag{12}$$

$\zeta\to\infty$ において式 (12) は式 (3) の $\mathcal{E}_0$ と一致する．また ζ を程々の値に抑えることで，$L(\Lambda)$ が滑らかな関数形となり，未知標本に対する頑健性を高めることができる [18], [21]．すなわち MCE 学習はベイズ誤り推定と直接的に結びついている．

$L(\Lambda)$ の最小化に関して，最急降下法などのバッチ的手法に加えて，Ω_N から 1 個の標本 (x_n,y_n) をランダム抽出する度に Λ を調整する適応的な学習方法も広く用いられている．その方法における Λ の調整機構は次式で与えられる．

$$\Lambda\leftarrow\Lambda-\varepsilon_n\nabla_\Lambda\ell\big(d_{y_n}(x_n,\Lambda)\big)\qquad(\varepsilon_n>0) \tag{13}$$

4.2 分類誤り数損失を通した MCE とブースティングとの関連性

2 値判別課題に限定した場合，MCE 学習における誤分類測度 $d_{y_n}(x_n,\Lambda)$ はマージンの正負を反転させたものに一致する．実際，$y_n=1$ の場合において，$d_1(x_n,\Lambda)=-g_1(x_n,\Lambda)+g_{-1}(x_n,\Lambda)=-y_nf(x_n,\Lambda)$ が導かれる（$y_n=-1$ の場合も同様）．よって式 (8) の評価汎関数 $C(F)$ は，c として式 (9) の $c_{\rm sig}$（平滑化された分類誤り数損失）を定めた場合，MCE 学習における式 (12)（および (11)）の評価関数 $L(\Lambda)$ と本質的に同一である．勾配法により，MarginBoost では F が探索され，MCE 学習では Λ が調整される．したがって，損失関数を平滑化された分類誤り数損失に定めることにより，MarginBoost における逐次的学習が，2 値判別課題に対する関数空間上の（関数を学習パラメータとした）MCE 学習に帰着し，MarginBoost がベイズ誤り推定を直接的に目指したものとなる．更に両者の関係を観測重みの観点で見てみよう．まず MCE 学習において，式 (13) におけるパラメータ修正項が

$$\nabla_\Lambda\ell\big(d_{y_n}(x_n,\Lambda)\big)=\ell'\big(d_{y_n}(x_n,\Lambda)\big)\cdot\nabla_\Lambda d_{y_n}(x_n,\Lambda) \tag{14}$$

となる．$\ell'\big(d_{y_n}(x_n,\Lambda)\big)$ は $d_{y_n}(x_n,\Lambda)$ が 0 の付近で大きな値となり，x_n が分類器の類境界付近に存在するときにパラメータ修正の程度が大きい．一方 $c=c_{\rm sig}$ による MarginBoost（図 2）において，w_n は $c_{\rm sig}$ の導関数に対応し，直前ステージまでの統合分類器のマージンが 0 の付近すなわち類境界付近に対して大きな重みをかけるものとなる．これはまさに MCE 学習におけるパラメータ調整機構と同等である．

5. Ensemble-based MCE

今までは 2 値判別課題を想定したが，実際的には多クラス分類課題がほとんどである．したがって，2 値判別手法を多クラス分類に拡張しなくてはならない．一般的に用いられる手段は，2 値判別課題を組み合わせる手法である [11], [16], [24], [25]．しかし容易にわかるとおり，多クラス分類におけるベイズ誤り推定との一貫性がなく，学習の最適性が保証されない．

そこで我々は，多クラス分類を行う弱分類器のアンサンブルに対して MCE 学習を適用し，多クラス分類誤りの最小化を直接的に目指す学習法である，アンサンブル型最小分類誤り（Ensemble-based MCE）学習法を提案する．2 値判別の MarginBoost では，$F_t=a_1f_1+\cdots+a_tf_t$ において t を増やしながら逐次的に f_t,a_t に関して $C(F_t)$ を最小化していた．Ensemble-based MCE はこれを多クラスに拡張した形式であり，以下のように構成される．まず，各ステージ t における多クラス弱分類器の判別関数である $h_{i,t}(x,\lambda_{i,t})$ $(i=1,...,U)$ を定める．ここで $\lambda_{i,t}$ は $h_{i,t}$ の学習パラメータ集合であり，$h_{i,t}$ は単純な構造の（少ないパラメータ数の）判別関数である．次に，ステージ t までに構成されるアンサンブル多クラス分類器の判別関数を

$$g_i^{(t)}(x,\Lambda_i^{(t)})=\sum_{\tau=1}^{t}a_{i,\tau}h_{i,\tau}(x,\lambda_{i,\tau})\ (i=1,...,U)\qquad \Lambda_i^{(t)}=\{\lambda_{i,1},...,\lambda_{i,t},a_{i,1},...,a_{i,t}\} \tag{15}$$

で定義する．$a_{i,t}$ は $h_{i,t}$ に対する分類器重みである．そして各 t において，判別関数 $g_i^{(t)}(x,\Lambda_i^{(t)})$ $(i=1,...,U)$ に対して MCE 学習を行う．これを $t=1,2...$ に関して増加的に実行する．この手続きを図 3 にまとめる．各 t において，$\lambda_{i,t},a_{i,t}$ 以外のパラメータ：$\lambda_{i,\tau},a_{i,\tau}$ $(\tau=1,...,t-1)$ は，直前ステージまでに求まったパラメータを初期値として調整できる．あるいはこれらを固定してもよく，この場合は前向きステージワイズの学習となる．また弱分類器 $\{h_{i,t}\}_{i=1}^{U}{}_{t=1}^{T}$ は，アンサンブル分類器に対する MCE 学習が逆伝搬する形で学習される．

Ensemble-based MCE のステージ t におけるパラメータ更新式は，式 (13)，(15) より，以下で与えられる(注9)．

$$\begin{aligned}\lambda_{y_n,t}&\leftarrow\lambda_{y_n,t}+\varepsilon_n\phi_na_{y_n,t}\nabla_{\lambda_{y_n,t}}h_{y_n,t}(x_n,\lambda_{y_n,t})\\ \lambda_{j_n,t}&\leftarrow\lambda_{j_n,t}-\varepsilon_n\phi_na_{j_n,t}\nabla_{\lambda_{j_n,t}}h_{j_n,t}(x_n,\lambda_{j_n,t})\end{aligned} \tag{16}$$

$$\begin{aligned}a_{y_n,t}&\leftarrow a_{y_n,t}+\varepsilon_n\phi_nh_{y_n,t}(x_n,\lambda_{y_n,t})\\ a_{j_n,t}&\leftarrow a_{j_n,t}-\varepsilon_n\phi_nh_{j_n,t}(x_n,\lambda_{j_n,t})\end{aligned} \tag{17}$$

$$\begin{aligned}\phi_n&=\ell'\big(D_{y_n}^{(t)}(x_n)\big)\\ D_{y_n}^{(t)}(x_n)&=-g_{y_n}^{(t)}(x_n,\Lambda_{y_n}^{(t)})+g_{j_n}^{(t)}(x_n,\Lambda_{j_n}^{(t)})\end{aligned} \tag{18}$$

ここで $j_n=\arg\max_{i\neq y_n}g_i^{(t)}(x_n,\Lambda_i^{(t)})$ であり，ϕ_n は $\{g_i^{(t)}(x_n,\Lambda_i^{(t)})\}_{i=1}^U$ に基づく誤分類測度 $D_{y_n}^{(t)}(x_n)$ に対する MCE 損失の勾配である．比較のために，t 番目弱分類器 $\{h_{i,t}\}_{i=1}^U$ のみの MCE 学習を考える．このときの更新式は，

$$\begin{aligned}\lambda_{y_n,t}&\leftarrow\lambda_{y_n,t}+\varepsilon_n\delta_n\nabla_{\lambda_{y_n,t}}h_{y_n,t}(x_n,\lambda_{y_n,t})\\ \lambda_{j_n,t}&\leftarrow\lambda_{j_n,t}-\varepsilon_n\delta_n\nabla_{\lambda_{j_n,t}}h_{j_n,t}(x_n,\lambda_{j_n,t})\end{aligned} \tag{19}$$

$$\begin{aligned}\delta_n&=\ell'\big(d_{y_n}(x_n)\big)\\ d_{y_n}(x_n)&=-h_{y_n,t}(x_n,\lambda_{y_n,t})+h_{j_n,t}(x_n,\lambda_{j_n,t})\end{aligned} \tag{20}$$

(注9)：ここでは式 (10) において $\eta\to\infty$ としており，また前向きステージワイズの学習を考える．

For $t = 1, 2, ..., T$ do:

1. Initialize $\lambda_{i,t}, a_{i,t}$ $(i = 1, ..., U)$.
2. Train the parameter set $\Lambda_i^{(t)}$ $(i = 1, ..., U)$ by applying MCE training to $g_i^{(t)}(x, \Lambda_i^{(t)})$ $(i = 1, ..., U)$.

Return U-class discriminat functions $g_1^{(T)}, ..., g_U^{(T)}$.

図 3 Ensemble-based MCE

となる．式 (16) と式 (19) との最も顕著な違いは，ϕ_n と δ_n である．弱分類器 $\{h_{i,t}\}_{i=1}^U$ のみの MCE 学習では，$\{h_{i,t}\}_{i=1}^U$ に関する類境界付近（δ_n が大きいところ）が重点的に学習される．これに対し，Ensemble-based MCE では，ステージ t までのアンサンブルである $\{g_i^{(t)}(x_n, \Lambda_i^{(t)})\}_{i=1}^U$ に基づく類境界付近（ϕ_n が大きいところ）に対して，大きな修正が行われる．すなわち各ステージにおいて，直前ステージまでの分類器が誤りを起こすサンプル群（外れ値を除く）に対して重点的に学習が行われ，相補的な学習と弱分類器の相関の小ささによる平滑効果という，ブースティングと同様の性質を有する．さらに，式 (17) より，弱分類器の正解クラスに対する判別関数 $h_{y_n,t}$ がより大きくなると（すなわち分類がより正確であるとき），分類器重み $a_{y_n,t}$ がより大きくなり，不正解クラスに対する判別関数 $h_{j_n,t}$ がより大きいとき（すなわち分類が不正確であるとき），分類器重み $a_{j_n,t}$ がより小さな値に修正される．これはブースティングにおける分類器重みの学習機構と同様の効果である．

そして重要なことは，Ensemble-based MCE が多クラス統合分類器に対して直接的に MCE 学習を適用していることであり，学習目的が，最終目標である多クラス分類のベイズ誤り状態を直接的に目指していることである．なお [5] でも MCE 基準に基づくアンサンブル分類器が提案されているが，そこでは $\{h_{i,t}\}$ が別途学習され $\{a_{i,t}\}$ のみがＭＣＥ学習によって求められるのに対し，Ensemble-based MCE では，$\{h_{i,t}\}$ と $\{a_{i,t}\}$ がともに MCE により（増加的に）学習される．

6. むすび

本稿では，統計的パターン認識において，複数の単純な分類器（弱分類器）を統合するアンサンブル型の MCE 学習法である，Ensemble-based MCE 学習法を提案した．この学習法は，多クラス分類におけるアンサンブル分類器のベイズ誤り状態を直接的に目指すものであり，しかもブースティングと同様の性質をもち，未知標本に対する高い頑健性が期待される．今後は様々な分類課題における実験により提案法の有効性を確認するとともに，提案法の理論的な解析を進めていく予定である．

謝辞 研究の機会を与えて下さった，情報通信研究機構 MASTER プロジェクト 音声コミュニケーショングループ 中村 哲 リーダー，清水 徹プロジェクトマネージャーに感謝します．本研究の一部は日本学術振興会 科学研究費補助金 基盤研究 (B)（課題番号：19300064）の援助により行われている．

文　献

[1] 石井健一郎，上田修功，前田英作，村瀬 洋，わかりやすいパターン認識，オーム社，東京，1998.
[2] R. O. Duda, P. E. Hart, and D. G. Stork（尾上守夫 監訳），パターン識別，新技術コミュニケーションズ，東京，2001.
[3] C. M. Bishop（元田 浩，栗田多喜夫，樋口知之，松本裕治，村田 昇 監訳），パターン認識と機械学習 上・下，シュプリンガー・ジャパン，東京，2007.
[4] L. Breiman, "Bagging predictors," Machine Learning 24, pp.123-140, 1996.
[5] 上田修功，"最小分類誤り基準に基づくニューラルネット識別機の最適線形統合法，" 信学論 (D-II), Vol. J82-DII, No. 3, pp.522-530, 1999.
[6] 上田修功，"アンサンブル学習の新展開，" 信学技報，PRMU2002-96, pp.31-36, 2002.
[7] 上田修功，"アンサンブル学習，" 情報処理学会論文誌，Vol.46, No.SIG15(CVIM 12), pp.11-20, 2005.
[8] R. E. Schapire, "The strength of weak learnability. " Machine Learning 5, pp.197-227, 1990.
[9] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of online learning and an application to boosting. " Journal of Computing System and Sciences, Vol 55, No. 1, pp.119-139, 1997.
[10] R. E. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions. " In Proceedings of the Eleventh Annual Conference on Computational Learning Theory, 1998.
[11] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," Ann. Statist., Vol. 28, No. 2, pp.337-407, 2000.
[12] L. Mason, J. Baxter, P.L. Bartlett, and M. Frean, "Functional gradient techniques for combining hypotheses." in Advances in Large Margin Classifiers, MIT Press, Cambridge, MA, 221-246, 2000.
[13] C. Domingo and O. Watanabe, "MadaBoost: A modification of AdaBoost." In Proc. 13th Conference on Computational Learning Theory, COLT' 00, pp.180-189, 2000.
[14] N. Sano, H. Suzuki, and M. Koda, "A robust boosting method for mislabeled data," J. Operations Research Society of Japan, Vol. 47, No. 3, pp.182-196, 2004.
[15] N. Sano, H. Suzuki, and M. Koda, "A robust boosting method using zero-one loss function : SNR Boost," 京都大学数理解析研究所講究録, Vol. 1351, pp.106-121, 2004.
[16] 金森敬文，畑埜晃平，渡辺 浩，ブースティング—学習アルゴリズムの設計技法—，森北出版，東京，2006.
[17] N. Cristianini and J.Shawe-Taylor（大北 剛 訳），サポートベクターマシン入門，共立出版，東京，2005.
[18] B.-H. Juang and S. Katagiri, "Discriminative training," J. Acoust. Soc. Jpn. (E), Vol. 13, No. 6, pp.333-339, 1992.
[19] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," IEEE Trans. Signal Processing, Vol. 40, No. 12, pp.3043-3054, 1992.
[20] S. Katagiri, B.-H. Juang, and C.-H. Lee, "Pattern recognition using a family of design algorithms based upon the generalized probabilistic descent method," Proc. IEEE. Vol. 86, No. 11, pp.2345-2373, 1998.
[21] E. McDermott and S. Katagiri, "Prototype-based MCE/GPD training for various speech units," Computer Speech and Language, Vol. 8, No. 4, pp.351-368, 1994.
[22] A. Biem, S. Katagiri, and B.-H. Juang, "Pattern recognition using discriminative feature extraction," IEEE Trans. Signal Processing, Vol. 45, No. 2, pp.500-504, 1997.
[23] H. Watanabe, T. Yamaguchi, and S. Katagiri, "Discriminative metric design for robust pattern recognition," IEEE Trans. Signal Processing, Vol. 45, No. 11, pp.2655-2662, 1997.
[24] T.G. Dietterich and G. Bakiri, "Solving multiclass learning problems via error-correcting output codes," Journal of Artificial Intelligence Research, No.2, pp.263-286, 1995.
[25] N. Miyake, T. Takiguchi, and Y. Ariki, "Noise detection and classification in speech signals with boosting," IEEE Statistical Signal Processing Workshop 2007, pp.778-782, 2007.