Logitboost による高精度早期認識法

藤野　知之†　　石黒　勝彦††

† 慶應義塾大学　〒 223–8522 神奈川県横浜市港北区日吉 3–14–1
†† 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: †fujino@thx.appi.keio.ac.jp, ††ishiguro.katsuhiko@lab.ntt.co.jp

あらまし　オンラインの時系列データ識別においては、多くの実用的な問題において「できるだけ早く正確に」識別を行うことが要求される。このような目的に基づく識別問題は「早期認識 (early classification)」と呼ばれる。近年、Adaboost に基づく早期認識モデルが提案され、その有効性が報告されている。本論文では、より確率的に精緻なモデルである Logitboost を利用した早期認識手法を提案する。提案法は、逐次ベイズ推定として自然に解釈可能であり、特に多クラス問題において顕著な性能向上が期待できる。実験の結果、手書き文字認識および動作認識の問題で良好な認識性能を示すことを確認した。

キーワード　早期認識、ブースティング、Logitboost

1. はじめに

時系列データの識別問題は機械学習における重要な問題の一つであり、オンライン音声認識 [1], オンライン手書き文字認識 [2]、動作認識 [3] など多くの応用分野がある。多くの時系列データ識別の研究では、隠れマルコフモデル (HMM) [4] に代表される生成モデルが利用されるが、一方で Support Vector Machine [5] や Conditional Random Field [6], Adaboost [7] のような識別モデルも適用可能である [8], [9]。

本論文であつかう問題は、時間長 T の時系列データに対し、それより早い時刻 $t \leqq T$ において、それまでに観測された早期時系列情報 $\{x(\tau)\}_{\tau=1}^{t}$ のみから高精度に時系列データのラベルを推定する「早期認識」という問題である。この問題は既存のモデルにおける目的、すなわち「時系列全体が与えられたとき、その目的関数を最大化する」という目標と異なるが、多くの実際的な場面において有用な課題である。早期認識の解法としては様々な手法が考えられるが、近年、Adaboost に基づく Earlyboost 法 [10] とその多クラス拡張 [11] が相次いで発表されており、その統計的なバックグラウンドと高い識別性能が報告されている。

本論文ではブースティングにもとづいた早期認識を Logistboost [12] の利用によって大幅に改良する手法を提案する。Logistboost は Logistic regression と最尤法の組み合わせによって Adaboost を改良した識別モデルである。本論文では Logitboost に逐次ベイズ推定モデルを適用することで、早期認識問題に応用する (Fig. 1) ことが可能であることを示す。また、手書き文字認識および動作認識実験において、顕著な認識率の向上を確認したので合わせて報告する。

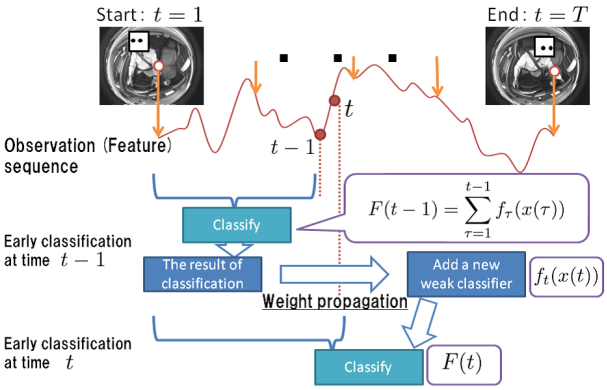

図 1: ブースティングによる早期認識の概要

2. 背　　景

2.1 Adaboost と Earlyboost

まず簡単に Adaboost を復習する。$x_i \in \mathbb{R}^d$ を学習データ、$y_i \in \{1, -1\}$ を対応するラベル情報とする。全データ数を N とする。Adaboost では強識別器 $F(x)$ を複数の弱識別器 $f_m(x) : \mathbb{R}^d \to \{-1, 1\}$ の線形結合により構成する:

$$F_M(x) = \sum_{m=1}^{M} c_m f_m(x). \qquad (1)$$

ここで c は弱識別器の importance weight であり、m は弱識別器のインデックスである。Adaboost の特徴は重み変数の伝播法にある。w_i を i 番目のデータに対する重みとすると、重み伝播法では m 番目の弱識別器で誤識別された (i.e. $y_i \neq f_m(x_i)$) データの重み w_i を増加させることでより良い強識別器を獲得する ($w_i \leftarrow w_i \exp(c_m)$)。

Adaboost は次の exponential loss [12] の最小化で説明される:

$$J(F) = \frac{1}{N}\sum_{i=1}^{N}[\exp(-y_i F(x_i))]\,. \qquad (2)$$

各 l 回目の繰り返し計算において、Adaboost アルゴリズムは $l-1$ 回の繰り返しで伝播されたデータの重み $\{w_i\}$ を利用して l 番目の弱識別器 $f_l(x)$ および importance weight c_l を最適化する。この時、Eq. (2) は Taylor 展開による近似を利用して下式に帰着する。

$$f_l(x) = \arg\min_{f(x)} \frac{1}{N}\sum_{i=1}^{N}[-w_i y_i f(x_i)]. \qquad (3)$$

Adaboost のアルゴリズムを Algorithm 1 に示す。ここで 1_S は、式 S が真なら 1, 偽なら 0 を返す関数である。

Algorithm 1 Standard Adaboost

Initialize $w_i = \frac{1}{N}, i = 1, \ldots, N$.
for $m = 1, 2, \ldots, M$ **do**
(1)Fit the weak classifier $f_m(x) \in \{-1, 1\}$ using weight w_i on the training dataset.
(2)$\epsilon_m = \mathrm{E}_w[1_{(y \neq f_m(x))}]$.
(3)$c_m = \log\left(\frac{(1-\epsilon_m)}{\epsilon_m}\right)$.
(4)$w_i \leftarrow w_i \exp[c_m 1_{y_i \neq f_m(x_i)}], i = 1, 2, \ldots, N$.
(5)$w_i \leftarrow \frac{w_i}{\sum_{i=1}^{N} w_i}$.
end for
Output sign$[F(x)]$, where $F(x) = \sum_{m=1}^{M} c_m f_m(x)$.

Earlyboost は以上の Adaboost を時系列データの早期認識へ応用した手法である [10]。相違点は、各学習時系列データ $\{x_i(t)\}_{i=1}^{N}$ のサンプル分布が、各時刻 $t = 1, 2, \ldots, T$ によって変化することである。Earlyboost では各時刻のサンプル分布が時刻ごとに独立であると仮定する。弱識別器 f_t は時刻 t ごとに独立であり、時刻 t に観測されたサンプルのみを入力を受け付ける。時刻の連続性はデータの重み $w_i(t)$ によってのみ担保される。従って強識別器は次のようになる。

$$F_T(x_i) = \sum_{t=1}^{T} c_t f_t(x_i(t)). \qquad (4)$$

Earlyboost の多クラス問題拡張は Earlyboost.MH として示されており [11]、またそこで統計的な背景も議論されているため、ここでは詳細を省略する.

2.2 Logitboost

Logitboost では弱識別器 $f(x) : \mathbb{R}^d \to \mathbb{R}$ が最尤法によって直接最適化される (重み変数が必要とされない)。強識別器は次式にあるように、重み変数を必要としない形で表現される。

$$F(x) = \sum_{m=1}^{M} \frac{1}{2} f_m(x) \qquad (5)$$

Logitboost では exponential loss の最小化に代えて "binomial likelihood function"(Eq. (6)) の最大化で学習を行う。このとき、あるデータ x_i のラベルが $y_i = 1$ となる確率は Eq. (7) で定義されている:

$$L(F) = \frac{1}{N}\sum_{i=1}^{N}[-\log(1 + \exp(-2y_i F(x_i)))]. \qquad (6)$$

$$p_i = \frac{1}{1 + \exp(-2F(x_i))}. \qquad (7)$$

尤度最大化のため、Logitboost は Newton 法を利用して以下の最小二乗誤差基準による推定式を得る。

$$f_l(x) = \arg\min_{f(x)} \sum_{i=1}^{N} w_i (z_i - f(x_i))^2, \qquad (8)$$

$$z_i = \frac{y_i^* - p_i}{p_i(1-p_i)}, \quad w_i = p_i(1-p_i), \qquad (9)$$

ここで $y_i^* = \frac{y_i+1}{2}$. Logitboost のアルゴリズムを Algorithm 2 に示す。

Algorithm 2 Standard Logitboost

Initialize $F(x) = 0$ and probability estimates $p_i = \frac{1}{2}$, $i = 1, 2, \ldots, N$
for $m = 1, 2, \ldots, M$ **do**
(1)Calculate the working response z_i and weight w_i by Eq. (9).
(2)Fit the weak classifier $f_m(x)$ using weighted least square as (8).
(3)$F(x) \leftarrow F(x) + \frac{1}{2} f_m(x)$ and $p_i \leftarrow \frac{1}{1+\exp(-2F(x_i))}$.
end for
Output the classifier sign$[F(x)]$

3. Logitboost に基づく早期認識法

3.1 2クラス問題に対する解法

本論文では sequential Logitboost、すなわち Logitboost による時系列データ早期認識法の応用を [11] で議論されたものと同様の方法によって実現する。時系列データ $x(t) = \{x_i(t)\}_{i=1}^{N}$ について各時刻の観測データが時刻ごとに独立であると仮定すると、Logitboost の枠組みをほぼそのまま応用可能となる。この場合、t 番目の弱識別器 f_t は時刻 t におけるサンプル集合 $x_{1;N}(t)$ にのみ適用する。

求める強識別器は次のように記述できる。

$$F(x_i) = \sum_{t=1}^{T} \frac{1}{2} f_t(x_i(t)) \qquad (10)$$

学習の目的は、上の強識別器を構成する時間依存の弱識別器 f_t を早期認識の枠組みで決定することである。そこで、時系列データの逐次ベイズ推定則に従って、時刻 t における $y_i^* = 1$ の事後確率を記述する。早期認識は

時刻 $t-1$ までの情報が得られた時に時刻 t で最大限の識別を行うという枠組みなので、逐次ベイズ推定はその目的にかなうはずであると考えられる。具体的に計算すると、

$$p(y_i^*=1|x(1:t)) = \frac{p(x_i(t)|y_i^*=1)p(y_i^*=1|x(1:t-1))}{\sum_{k=\{0,1\}} p(x_i(t)|y_i^*=k)p(y_i^*=k|x(1:t-1))} \quad (11)$$

を得る。以下、$x(1:t)=\{x(\tau)\}_{\tau=1}^{t}$ とする. シグモイド関数を $\sigma(a)=\frac{1}{1+\exp(-a)}$ で表すと、事後確率は次のようにパラメタライズされる。

$$p(y_i^*=1|x(1:t)) = \sigma(a(t)). \quad (12)$$

ここでパラメータに着目すると以下のような分解を得る。

$$a(t) = \ln\frac{p(x_i(t)|y_i^*=1)}{p(x_i(t)|y_i^*=0)} + \ln\frac{p(y_i^*=1|x(1:t-1))}{p(y_i^*=0|x(1:t-1))}. \quad (13)$$

Eq. (13) の第一項は $x_i(t)$ に関する尤度比の対数になっている。また第二項は時刻 $t-1$ における早期認識事後分布の対数比であり、時刻 t における事前分布の対数比として利用していると解釈できる。ここで、第一項を $f_t(x_i(t))$ と表現し、$t=1$ において第二項を 0、すなわち観測値を受信する前は 2 クラス間に対する知識がまったく等価であるという自然な仮定を認めると、以下の再帰式を得る。

$$\begin{aligned} a(t) &= f_t(x(t)) + a(t-1) \\ &= f_t(x(t)) + \sum_{s=1}^{t-1} f_s(x(s)) \end{aligned} \quad (14)$$

このように、逐次ベイズ推定の枠組みで早期認識を実現しようとすることで、弱識別器の additive model が自然と導出される。

時刻 t までの binomial log-likelihood は以下の式で表現される:

$$L(f_{1:t}) = \frac{1}{N}\sum_{i=1}^{N}\log\frac{1}{1+\exp\left(-2y_i\left(\sum_{s=1}^{t-1} f_s(x_i(s)) + f_t(x_i(t))\right)\right)}. \quad (15)$$

時刻 t における prior を p_i と書くと、Newton 法 $F(x) \leftarrow F(x) - H^{-1}(x)s(x)$ で必要となる偏微分は以下のようになる:

$$\begin{aligned} s(x(1:t)) &= \left.\frac{\partial L}{\partial f_t(x(t))}\right|_{f_t(x(t))=0} \\ &= \frac{2}{N}\sum_{i=1}^{N}(y_i^* - p_i), \end{aligned} \quad (16)$$

$$\begin{aligned} H(x(1:t)) &= \left.\frac{\partial^2 L}{\partial f_t(x(t))^2}\right|_{f_t(x(t))=0} \\ &= -\frac{4}{N}\sum_{i=1}^{N} p_i(1-p_i). \end{aligned} \quad (17)$$

各弱識別器 f_t は Eq. (8) と同じ最小二乗基準で求める. 早期認識 Logitboost 法のアルゴリズムを Algorithm 3 に示す。

Algorithm 3 Binary Logitboost for early classification

Initialize $F(x)=0$ and probability estimates $p_i=\frac{1}{2}$, $i=1,\ldots,N$.

for $t=1,2,\ldots,T$ **do**

(1)Calculate the working response z_i and weight w_i by Eq. (9).

(2)Fit the weak classifier $f_t(x(t))$ using weighted least square as (8).

(3)$F(x(1:t)) \leftarrow F(x(1:t-1)) + \frac{1}{2}f_t(x(t))$ and $p_i \leftarrow \frac{1}{1+\exp(-2F(x_i(1:t)))}$.

end for

Output the classifier $\mathrm{sign}[F(x(1:\tau))]$ at any time $1 \leqq \tau \leqq T$

3.2 多クラス問題に対する解法

多クラス Logitboost [12] では以下の尤度関数の最大化を行う:

$$L(y^*,p) = \sum_{k=1}^{K} y_k^* \log p_k, \quad (18)$$

$$p_{k,i} = \frac{\exp F_k(x_i)}{\sum_{j=1}^{K}\exp F_j(x_i)}. \quad (19)$$

ここで $k=1,\ldots,K$ はクラスのインデックス、K は総クラス数、$y_k^*=\{0,1\}$ はクラス k に対するラベル、そして F_k クラス k に対する識別器をそれぞれ表す。sequential multi-class Logitboost、すなわち多クラス早期認識法も 2 クラスの場合と同様に導出可能である。

Base class K を任意に選んだ上で、ベイズ則に従って尤度 Eq. (18) を以下のように書き換える:

$$L(g_{1:K,1:t}) = \frac{1}{N}\sum_{i=1}^{N}\Big[\sum_{k=1}^{K-1} y_{k,i}^*\left(\sum_{s=1}^{t-1} g_{k,s}(x_i(s)) + g_{k,t}(x_i(t))\right) - \log\left(1+\sum_{k=1}^{K-1}\exp\left(\sum_{s=1}^{t-1} g_{k,s}(x_i(s)) + g_{k,t}(x_i(t))\right)\right)\Big]. \quad (20)$$

ここで $g_{k,t}(x_i(t)) = \log p_{k,i} - \log p_{K,i}$; また $g_{K,t}$ は全ての t についてゼロとする。上式を偏微分し、以下の 2 式を得る:

$$s_k(x(1:t)) = \frac{1}{N}\sum_{i=1}^{N}(y^*_{k,i} - p_{k,i}), \tag{21}$$

$$H_{j,k}(x(1:t)) = -\frac{1}{N}\sum_{i=1}^{N} p_{j,i}(\delta_{j,k} - p_{k,i}). \tag{22}$$

ただし $j, k = 1, \ldots, J-1$. ヘッセ行列を対角行列で近似することで以下の更新則を求めることができる。

$$g_{k,t}(x_i(t)) \leftarrow \frac{y^*_{k,i} - p_{k,i}}{p_{k,i}(1-p_{k,i})}. \tag{23}$$

弱識別器は base class$K$ を除いて $w_i = p_{k,i}(1-p_{k,i})$ を重みとする $g_{k,t}(x_i(t))$ の重み付き二乗誤差基準で求まる。しかし、結果の対称性を担保するため、以下の更新則を採用するのが一般的である。

$$f_{k,t}(x_i(t)) = \frac{K-1}{K}\left(g_{k,t}(x_i(t)) - \frac{1}{K}\sum_{k=1}^{K} g_{k,t}(x_i(t))\right) \tag{24}$$

ここで全ての k について $g_{k,t},\ k = 1, \ldots, K$ は Eq. (23) に従って計算する. Logitboost による多クラス早期認識手法のアルゴリズムを Algorithm 4 にまとめる。

Algorithm 4 Multi-class Logitboost for early classification

Initialize $F_k(x) = 0$ and probability estimates $p_{k,i} = \frac{1}{K}$, $k = 1, \ldots, K, i = 1, \ldots, N$.
for t=1,2,..., T **do**
(1)Calculate the working response $g_{k,t}$ by eq. (23) and weight $w_i = p_{k,i}(1-p_{k,i})$.
(2)Fit the weak classifier $f_{k,t}(x(t))$ into $g_{k,t}(x(t))$ using weighted least squares (8).
(3)Set $f_{k,t} \leftarrow \frac{K-1}{K}(f_{k,t}(x(t)) - \frac{1}{K}\sum_{k=1}^{K} f_{k,t}(x(t)))$
(4)Update $F_k(x(1:t)) = F_k(x(1:t-1)) + f_{k,t}(x(t))$ and $p_{k,i} \leftarrow \frac{\exp(F_k(x_i(1:t)))}{\sum_{j=1}^{K}\exp(F_j(x_i(1:t)))}$
end for
Output the classifier $\arg\max_k[F_k(x(1:\tau))]$ at any time $1 \leqq \tau \leqq T$

4. 実　　験

4.1 実験設定

本論文では 2 種類の実験を行った (Fig. 2). 最初の実験はオンライン手書き文字認識である。“Kuchibue” データベースよりアルファベットとひらがなのデータセットを収集して識別実験を行った。一つの時系列データが一文字の軌跡に相当する。各時系列データの長さは全て単純な線形補間により $T = 50$ に整形した。観測特徴 $x_i(t)$ は各時刻におけるスタイラスペンの軌跡および速度である ($d = 4$). アルファベットデータはクラス数 $K = 52$(大文字、小文字を区別)、時系列データの数 $N = 14000$ である。一方、ひらがなデータはクラス数 $K = 83$, 時系列データの数 $N = 52500$ である。識別性能は 10 交差検定によって計算した。

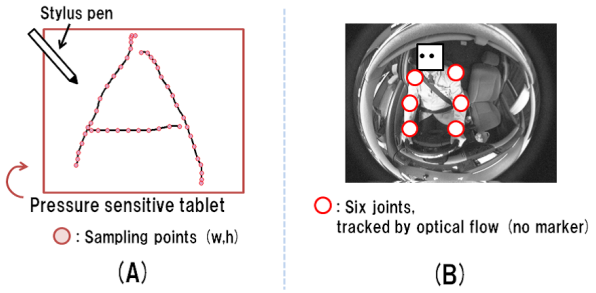

図 2: 実験データの例。 (A) 手書き文字認識はアルファベットとひらがなを用いて、できる限り早く書いている文字を推定する (B) 運転動作認識では optical flow によって関節位置を追跡し、実際に車内機器を操作する前に動作を識別する

二番目の実験は運転中の動作認識問題である。ドライブシミュレーターにカメラを設置し、ドライバーの行動を 60fps で撮影した。実験には 7 名の被験者が参加し、それぞれ 30 回の運転シミュレーションを実施した。ドライバーの左右の手首、肘及び肩の計 6 関節の 2 次元画像上の位置および速度を optical flow を利用して推定・追跡しそれを観測特徴として利用した。すなわち $x_i(t)$ は $d = 24$ 次元ベクトルとなる。認識する動作は “エアコンを操作する”、“バックミラーを触る” など、$K = 12$ 種類の車内の機器操作を行う動作を準備した。これらはステアリングから手を離す必要があるため、手を離した瞬間を動作の開始点とする。全ての時刻フレームは人手によりラベルづけされ、また一区切りの動作ごとに分節化した。各被験者ごとの時系列データ数は一定ではないが、およそ被験者 1 名に対して $N = 650$ のシーケンスが採取された。各被験者ごとに 6 交差検定によって識別性能を評価し、最後に 7 人の被験者の成績の単純な平均で最終的な数値評価を行う。

4.2 実験結果

実験では提案する multi-class sequential Logitboost と Adaboost に基づく多クラス早期認識法 Earlyboost.MH [11] を比較した。弱識別器としては双方とも decision stump を利用した。アルファベットに対する平均誤識別率を Fig. 3 に示す。縦軸は誤識別率でありは横軸は入力されたテスト時系列シーケンスの長さ t である。同様にひらがなの誤認識率を Fig. 4 に、運転動作の誤認識率を Fig. 5 にそれぞれ示す。赤実線が提案法であり、青破線が Earlyboost.MH の結果である。

全ての実験において、Logitboost に基づく提案法が Adaboost に基づく Earlyboost.MH よりも早期認識の観点で良い性能、すなわち、より短い入力シーケンスで

表 1: 実験データセット

Task	アルファベット	ひらがな	運転動作
クラス数 K	52	83	12
時系列データ数 N	14000	52500	被験者毎に $\simeq 650$
特徴量の次元 d	4	4	24

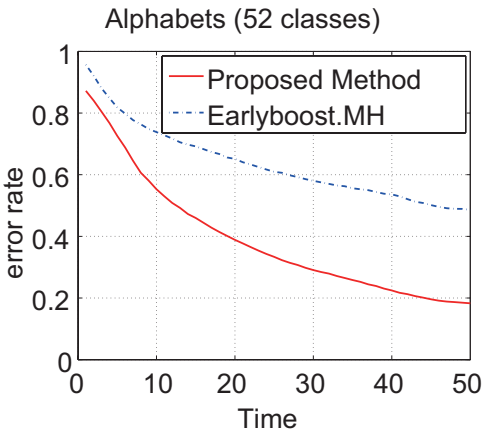

図 3: アルファベット文字認識実験の結果。赤実線は提案法の平均誤識別率を、青破線は Earlyboost.MH 法の平均誤識別率を示す。

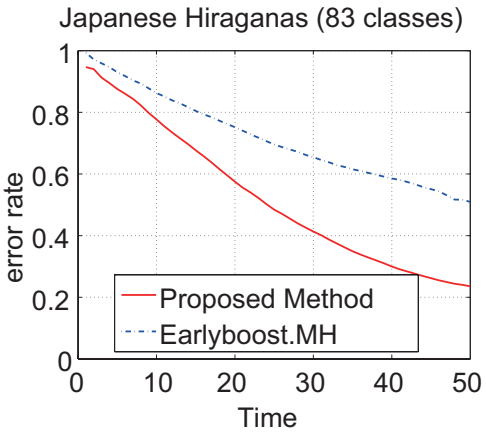

図 4: ひらがな認識実験の結果。赤実線は提案法の平均誤識別率を、青破線は Earlyboost.MH 法の平均誤識別率を示す。

より低い誤識別率を達成した。一般的に Logitboost は Adaboost よりも良い識別性能を示すことが知られている。その理由の一つは最適化法の違いがある。また、この早期認識の枠組みでは学習可能なパラメータ数の違いも影響していると思われる。Earlyboost.MH では弱識別器が $f(x) : \mathbb{R}^d \to \{-1, 1\}$ に制約されており、さらにそれらの間のバランスは T 個の重み $\{c_t\}_{t=1}^T$ のみで実現されている。一方、提案する multi-class sequential Logitboost では弱識別器は $f(x) : \mathbb{R}^d \to \mathbb{R}$ であり、これを直接最適化している。これは Earlyboost.MH において $2 \times T \times K$ 個のパラメータを最適化しているのと同じである。従って Logitboost はより良い fit を得ることができる。

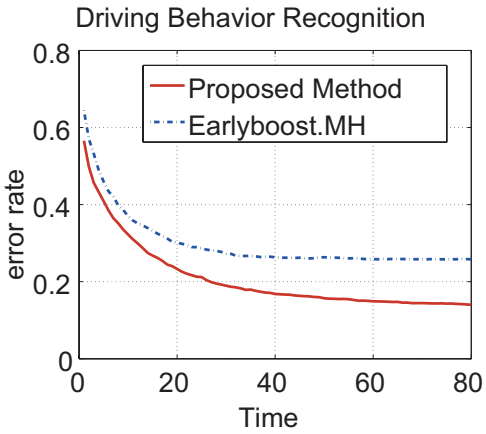

図 5: 運転中機器操作動作の認識実験の結果。赤実線は提案法の平均誤識別率を、青破線は Earlyboost.MH 法の平均誤識別率を示す。

5. 結　　論

本論文では、Logitboost を利用した時系列データの早期認識手法を提案した。我々は 2 クラスおよび多クラス問題それぞれに対する定式化をベイズ推定の枠組みで導出した。さらに実験によって Adaboost に基づく既存の早期認識手法よりも顕著に良い識別性能を達成可能であることを示した。

文　　献

[1] T. Hori, C. Hori and Y. Minami: "Fast on-the-fly composition for weighted finite-state transducers in 1.8 million-word vocabulary continuous speech recognition", Proceedings of the 8th International Conference on Spoken Language Processing (Interspeech), Vol. 1, pp. 289–292 (2004).

[2] C. Bahlmann and H. Burkhardt: "The writer independent online handwriting recognition system frog on hand and cluster generative statistical dynamic time warping", IEEE Transactions on Pattern Analysis and Machine Intelligence, **26**, 3, pp. 299–310 (2004).

[3] Y. A. Sheikh, A. Datta and T. Kanade: "On the sustained tracking of human motion", Proceedings of 8th IEEE International Conference on Automatic Face and Gesture Recognition (FG) (2008).

[4] L. R. Rabiner: "A tutorial on hidden Markov models and selected applications in speech recognition", Proceedings of the IEEE, **77**, 2, pp. 257–286 (1989).

[5] V. N. Vapnik: "The Nature of Statistical Learning Theory", Springer (1995).

[6] J. Lafferty, A. McCallum and F. Pereira: "Conditional random field: Probabilistic models for segmenting and labeling sequence data", Proceedings of 18th International Conference on Machine Learning, pp. 282–289 (2001).

[7] Y. Freund and R. E. Schapire: "A decision-theoretic generalization of on-line learning and an application to boosting", Journal of Computing Systems and Science, **55**, 1, pp. 119–139 (1997).

[8] K. J. Kim: "Financial time series forecasting using support vector machines", Neurocomputing, **55**, pp. 307–319 (2003).

[9] A. Gunawardana, M. Mahajan, A. Acero and J. C. Platt: "Hidden conditional random fields for phone classification", Proceedings of Interspeech (2005).

[10] S. Uchida and K. Amamoto: "Early recognition of sequential patterns by classifier combination", Proceedings of the 19th International Conference on Pattern Recognition (ICPR) (2008).

[11] K. Ishiguro, H. Sawada and H. Sakano: "Multi-class boosting for early classification of sequences", Proceedings of the Twenty-first British Machine Vision Conference (BMVC), pp. 24.1–24.10 (2010).

[12] J. Friedman, T. Hastie and R. Tibshirani: "Additive logistic regression: A statistical view of boosting (with discussion)", Annals of Statistics, **28**, 2, pp. 337–407 (2000).