統計数理(2010)
第 58 巻 第 2 号 157–166

特集「統計的機械学習」
[研究ノート]

2クラス判別器の組み合せによる多クラス判別統計モデルとパラメータ推定

池田 思朗†

(受付 2009 年 12 月 25 日；改訂 2010 年 1 月 19 日；採択 3 月 15 日)

要　　旨

機械学習の分野では，2 クラス判別器を組み合せて多クラスの判別器を作る様々な試みがなされている．Error Correcting Output Codes (ECOC) と呼ばれる方法が頻繁に用いられているが，それぞれの 2 クラス判別器が軟判定を返す場合には ECOC の他にも Bradley-Terry (BT) モデルを用いる方法が提案されている．本稿ではこの BT モデルを用いる組み合せ法を統計モデルとして考え，既存の方法の改善法のひとつを示す．

キーワード： Bradley-Terry モデル，多クラス判別，最尤推定．

1. はじめに

機械学習の分野では以前より 2 クラス判別器を組み合せ，多クラス判別を行なう試みがなされている(Dietterich and Bakiri, 1995)．2 クラス判別器を組み合せる理由のひとつは，Support Vector Machine (SVM) (Boser et al., 1992) のように性能の良い 2 クラス判別のための手法が提案されていることである．ただし，多クラスへの拡張は簡単ではない．また，別の理由として，クラスの数が非常に多い場合，全てのデータを用いてパラメータ推定を行なうよりも，2 クラスに限定して学習する 2 クラス判別器を数多く構成する方が効率が良いことも挙げられる．

2 クラス判別器を組み合せる手法として提案され，広く用いられているのは ECOC による多クラス判別器の構成法(Dietterich and Bakiri, 1995; Allwein et al., 2000)である．この手法は簡単であり，さまざまな拡張法が報告されている．一方，これとは別に BT モデル(Bradley and Terry, 1952)を用いた方法が提案されている(Hastie and Tibshirani, 1998; Zadrozny, 2002)．さらに Huang らは BT モデルを拡張し，generalized Bradley-Terry (GBT) モデルを提案した(Huang et al., 2006)．様々な提案はされているが，多くの手法に関する解析は，数値実験を通じた結果の比較にとどまっている．あるいは，個々の 2 クラス判別器の出力にある種の統計性を仮定した議論を行なっている場合があるが，その仮定を正当化することは困難である．

本稿では特に GBT モデルに注目し，統計モデルとしてとらえ，パラメータの推定法を提案する．その過程で ECOC モデルと GBT モデルを用いる方法の関係についても議論する．

2. 2 クラス判別器と ECOC 法

本稿では K-クラスの判別問題を考える $(K>2)$．特徴ベクトルを $\boldsymbol{x}\in\mathcal{X}$，クラスラベルを $y\in\{1,\ldots,K\}$ とおく．多クラス判別の問題は $\boldsymbol{x}$ からクラスラベル y を予測する判別器を構成

† 統計数理研究所：〒190–8562 東京都立川市緑町 10–3

することである．ただし推定のために N 個のサンプル $(\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_N,y_N)$ が与えられているとする．その際，2 クラス判別器を組み合せて多クラス判別器を構成することを考える．まず，2 クラス判別器を定義する．

2.1 2 クラス判別器

R 組のクラスラベルの部分集合 C_r^+，C_r^- を次のように定義する．

$$C_r^+,C_r^-\subset\{1,\ldots,K\},\ C_r^+,C_r^-\neq\emptyset,\ C_r^+\cap C_r^-=\emptyset,\quad C_r=C_r^+\cup C_r^-,\quad r=1,\ldots,R.$$

それぞれの 2 クラス判別器は $\boldsymbol{x}$ が C_r^+ のクラスか C_r^- のクラスに属するかを判別し，軟判定（ここでの軟判定は，各クラスに属する確率値を返す判定のことを示す．これに対してひとつのクラスだけを選ぶ判定は硬判定と呼ぶ）の結果を返す．かりに $|C_r^+|=|C_r^-|=1$ であるならば “1-vs.-1” 判別器と呼ばれ，$|C_r^+|=1$，$|C_r^{-1}|=K-1$ ならば “1-vs.-the rest” 判別器と呼ばれる．それぞれの判別器は軟判定を行なうとする．すなわち r 番目の 2 クラス判別器は $q_r(\boldsymbol{x})$ を返す．これは $p(y\in C_r^+|y\in C_r,\boldsymbol{x})$ を推論したものである．なお，$1-q_r(\boldsymbol{x})$ は $p(y\in C_r^-|y\in C_r,\boldsymbol{x})$ を推論したものに相当する．以下では，各 2 クラス判別器がパラメータ $\boldsymbol{\eta}_r=(\eta_{r1},\ldots,\eta_{rp})^T$ を持ち，$\boldsymbol{\eta}_r$ で微分可能とし $q_r(\boldsymbol{x};\boldsymbol{\eta}_r)$ と書く．さらに次の性質を仮定する．

$$0<q_r(\boldsymbol{x};\boldsymbol{\eta}_r)<1,\quad \forall\boldsymbol{x}\in\mathcal{X}.$$

例えば古くから広く用いられているロジスティック回帰（Hosmer and Lemeshow, 2000）はこの条件を満たす．

2.2 ECOC に基づく多クラス判別器と統計モデル

ECOC に基づく多クラス判別器の構成は広く用いられている手法である．以下で簡単に説明する．まず，行列 $M\in\{-1,0,+1\}^{K\times R}$ を以下のように定義する（Allwein et al., 2000）．

$$m_{kr}=\begin{cases}+1, & k\in C_r^+\\ -1, & k\in C_r^-\\ 0, & k\notin C_r.\end{cases}\tag{2.1}$$

これが符号行列と呼ばれたことが ECOC の名前の由来である．$\boldsymbol{x}$ が観測されたとき，2 クラス判別器 $q_r(\boldsymbol{x};\boldsymbol{\eta}_r)$，$(r=1,\ldots,R)$ の出力を組み合せ，以下のように多クラスのクラスラベル $y\in\{1,\ldots,K\}$ の判別を行なう．

$$\hat{y}=\operatorname*{argmax}_k\left[\sum_{r=1}^R m_{kr}q_r(\boldsymbol{x};\boldsymbol{\eta}_r)\right].\tag{2.2}$$

(2.2) 式に定義された ECOC の判別方法を拡張して確率 $p(y|\boldsymbol{x})$ の推論を行なうことを考えよう．例えば任意の正値を取る単調増加関数 $g(\cdot)$ をとり，次のように確率を定義すれば

$$p_{ECOC}(y|\boldsymbol{x};\{\boldsymbol{\eta}_r\})=\frac{g\left(\sum_{r=1}^R m_{yr}q_r(\boldsymbol{x};\boldsymbol{\eta}_r)\right)}{\sum_{k=1}^K g\left(\sum_{r=1}^R m_{kr}q_r(\boldsymbol{x};\boldsymbol{\eta}_r)\right)},\ y\in\{1,\ldots,K\},\ \{\boldsymbol{\eta}_r\}=\{\boldsymbol{\eta}_1,\ldots,\boldsymbol{\eta}_R\}.\tag{2.3}$$

判別結果は $\hat{y}=\operatorname{argmax}_k\ p_{ECOC}(k|\boldsymbol{x};\{\boldsymbol{\eta}_r\})$ となる．関数 $g(\cdot)$ としては様々なものを考えることができるが，特に $\exp(\cdot)$ ととれば，この関数は多項ロジスティック回帰モデルとなる．同様な解釈に基づく ECOC の拡張は既に提案されており，ここでは深くは触れない（Crammer and Singer, 2002）．

3. GBT モデル

1-vs.-1 判別器を組み合せて多クラス判別器を構成する方法として BT モデルを用いる方法が

Hastie and Tibshirani (1998)によって提案された．まず，BT モデル(Bradley and Terry, 1952)について説明し，それを用いてどのように多クラス判別器を構成するかを説明する．

3.1 BT モデル

多数のプレーヤーが 1 対 1 の試合を数多く行なったとき，それぞれのプレーヤーのランク付けをする問題へのひとつのアプローチとして BT モデルの説明をする．まず K 人のプレーヤーがいると仮定する．それぞれのプレーヤーは別のプレーヤーと 1 対 1 で何らかの勝負をする．テニスのようなスポーツでもチェスのようなボードゲームでもよい．そして対戦の組み合せを変えた勝負の結果が数多く得られたとする．

BT モデルでは，プレーヤー i は p_i という正の値を持っていると仮定し，あるプレーヤー k が別のプレーヤー l に次の確率で勝つと仮定する．

$$P(k \text{ beats } l) = \frac{p_k}{p_k + p_l},\ p_{k'} > 0\ \ (k' = 1, \ldots, K).$$

これは全ての p_k を定数倍しても変わらないので $\sum_k p_k = 1$ と規格化する．何回かのゲームの結果が得られたとき，プレーヤー k が別のプレーヤー l に勝った割り合いを次のように r_{kl} と定義する．

$$n_{kl} = \#(\text{games between } k \text{ and } l),\ r_{kl} = \frac{\#(k \text{ beats } l)}{n_{kl}}. \tag{3.1}$$

簡単のために引き分けはないものとしよう．すなわち $r_{kl} + r_{lk} = 1$ である．次の対数尤度を考える．

$$F(\boldsymbol{p}) = \sum_{k<l} n_{kl} \left(r_{kl} \ln \frac{p_k}{p_k + p_l} + (1 - r_{kl}) \ln \frac{p_l}{p_k + p_l} \right). \tag{3.2}$$

パラメータ $\boldsymbol{p} = (p_1, \ldots, p_K)$ に関する最尤推定は以下のように定義できる．

$$\hat{\boldsymbol{p}} = \operatorname*{argmax}_{\boldsymbol{p}} F(\boldsymbol{p}), \quad \text{subj. to } \sum_{k=1}^{K} \hat{p}_k = 1,\ \hat{p}_k > 0.$$

この結果得られる $\hat{\boldsymbol{p}}$ の各成分は各プレーヤーの強さを表していると考え，各プレーヤーのランク付けを行なう． $F(\boldsymbol{p})$ の最大化の計算は一般にそれほど難しくない(Hunter, 2004)．

3.2 BT モデルによる多クラス判別器の構成

2 クラス判別器を組み合せる際に BT モデルを用いるには BT モデルにおける(3.1)式の r_{kl} を 1-vs.-1 判別器の出力 $q_r(\boldsymbol{x}; \boldsymbol{\eta}_r)$ に変更し，$n_{kl} = 1$ と置く，すなわち(3.2)式の尤度に対応する量は以下のように定義される．

$$F_{BT}(\boldsymbol{p}, \boldsymbol{x}, \{\boldsymbol{\eta}_r\}) = \sum_{r=1}^{R} \left(q_r(\boldsymbol{x}; \boldsymbol{\eta}_r) \ln \frac{p_{k_r}}{p_{k_r} + p_{l_r}} + (1 - q_r(\boldsymbol{x}; \boldsymbol{\eta}_r)) \ln \frac{p_{l_r}}{p_{k_r} + p_{l_r}} \right), \tag{3.3}$$

ただし k_r はそれぞれの C_r^+ のクラス，l_r は C_r^- のクラスである．BT モデルを用いた多クラスの判別器では $F_{BT}(\boldsymbol{p}, \boldsymbol{x}, \{\boldsymbol{\eta}_r\})$ を最大にする $\hat{p}_k$ が $p(y|\boldsymbol{x})$ の推定値だと考える．

$$p_{BT}(y|\boldsymbol{x}; \{\boldsymbol{\eta}_r\}) = \hat{p}_y,\ \hat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_K) = \operatorname*{argmax}_{\boldsymbol{p}} F_{BT}(\boldsymbol{p}, \boldsymbol{x}, \{\boldsymbol{\eta}_r\}), \quad \text{subj. to } \sum_{k=1}^{K} \hat{p}_k = 1,\ \hat{p}_k > 0.$$

この結果を用いて，推論や判別を行なう．

3.3 GBT モデルによる多クラス判別器の構成と統計モデル

上記の手法は 1-vs.-1 判別器のみならず, 任意の判別器の組み合せに拡張可能である(Zadrozny, 2002; Huang et al., 2006). その際, (3.3)式の関数 $F_{BT}(\boldsymbol{p},\boldsymbol{x},\{\boldsymbol{\eta}_r\})$ を次の $F_{GBT}(\boldsymbol{p},\boldsymbol{x},\{\boldsymbol{\eta}_r\})$ へ拡張する.

$$F_{GBT}(\boldsymbol{p},\boldsymbol{x},\{\boldsymbol{\eta}_r\})=\sum_{r=1}^{R}\left(q_r(\boldsymbol{x};\boldsymbol{\eta}_r)\ln\frac{\sum_{k\in C_r^+}p_k}{\sum_{l\in C_r}p_l}+\bigl(1-q_r(\boldsymbol{x};\boldsymbol{\eta}_r)\bigr)\ln\frac{\sum_{k\in C_r^-}p_k}{\sum_{l\in C_r}p_l}\right). \tag{3.4}$$

$$p_{GBT}(y|\boldsymbol{x};\{\boldsymbol{\eta}_r\})=\hat{p}_y,\ \hat{\boldsymbol{p}}=(\hat{p}_1,\ldots,\hat{p}_K)=\operatorname*{argmax}_{\boldsymbol{p}}F_{GBT}(\boldsymbol{p},\boldsymbol{x},\{\boldsymbol{\eta}_r\}),$$

$$\text{subj. to }\sum_{k=1}^{K}\hat{p}_k=1,\ \hat{p}_k>0.$$

クラス集合 $\{C_r^+\}$ と $\{C_r^-\}$ の選び方によっては, F_{GBT} を最大にする $\boldsymbol{p}$ が複数ある可能性がある. F_{GBT} が唯一の最大値を持つ十分条件は Huang らによって調べられている(Huang et al., 2006). その中で, **全ての 1-vs.-1 判別器を組み合せる場合** と**全ての 1-vs.-the rest 判別器を組み合せる場合** には唯一の極大点を持ち, それが最大値を与えることが示された. 本稿では簡単のため BT モデルの場合を扱うが, GBT モデルが唯一の極大点を持つ場合への拡張は簡単である.

BT モデル, あるいは GBT モデルを用いる場合, どのようにして $F_{GBT}(\boldsymbol{p},\boldsymbol{x},\{\boldsymbol{\eta}_r\})$ の最大値を計算するかが問題となる. 複数のアルゴリズムが提案されている(Hunter, 2004; Huang et al., 2006). どのアルゴリズムを用いてもよいが, 本稿ではこのうちのひとつで GBT モデルへの拡張が容易である MM (minorization-maximization) アルゴリズム(Hunter, 2004)を用いた.

3.4 GBT モデルと ECOC 法の関係について

BT モデルでは(3.2)式が対数尤度関数として導かれた. この関数の最大化によって得られるパラメータは最尤推定量であり, 最尤推定の観点から(3.2)式の関数は重要である. しかしながら, 多クラス判別器を作る際の(3.4)式の関数は(3.2)式の関数と類似しているが, 統計的観点からはこの関数を用いる積極的な理由は見出せない.

第一に $q_r(\boldsymbol{x};\boldsymbol{\eta}_r)$ は試合の結果ではない. また, $q_r(\boldsymbol{x};\boldsymbol{\eta}_r)$ は学習の段階では $p(y\in C_r^+|y\in C_r,\boldsymbol{x})$ の推論として得られたものである. にもかかわらず, 判別の際には $y\notin C_r$ だと思われる $\boldsymbol{x}$ に対する $q_r(\boldsymbol{x};\boldsymbol{\eta}_r)$ の値を用いるのである. したがって, GBT モデルから導かれた尤度関数 F_{GBT} を多クラス判別に用いる際には, これは単に $\{q_r\}$ と $\boldsymbol{p}$ とを結びつけるための関数だととらえるべきだろう. この際 $\boldsymbol{p}$ の関数 F に要求される性質は $\boldsymbol{p}$ に関して唯一の解を持ち, それが容易に計算できることである.

このような関数は様々なものがある. 例えば, (2.2) 式の関数 $g(\cdot)$ を用いて,

$$F_{ECOC}(\boldsymbol{p},\boldsymbol{x},\{\boldsymbol{\eta}_r\})=\sum_k g\left(\sum_{r=1}^{R}m_{kr}q_r(\boldsymbol{x};\boldsymbol{\eta}_r)\right)\ln p_k$$

とおき,

$$p_{ECOC}(y|\boldsymbol{x};\{\boldsymbol{\eta}_r\})=\hat{p}_y,\ \hat{\boldsymbol{p}}=(\hat{p}_1,\ldots,\hat{p}_K)=\operatorname*{argmax}_{\boldsymbol{p}}F_{ECOC}(\boldsymbol{p},\boldsymbol{x},\{\boldsymbol{\eta}_r\}),$$

$$\text{subj. to }\sum_{k=1}^{K}\hat{p}_k=1,\ \hat{p}_k>0$$

という問題を考えると, この結果は(2.3)式の $p_{ECOC}(y|\boldsymbol{x};\{\boldsymbol{\eta}_r\})$ と一致していることが簡単にわかる. BT モデルと GBT モデルは ECOC とは別のモデルだが, その差は q_r と $\boldsymbol{p}$ の間を結ぶ非線形関数の差が問題であり, 上のように定式化できる.

他にも F の拡張は可能である. (3.3)式は Kullback-Leibler ダイバージェンスを用いて表現できる．Ikeda (2006) ではこれを甘利によって定義された α ダイバージェンス (Amari, 1985) へ拡張した場合についての報告がされている.

4. BT モデルのパラメータ推定

4.1 パラメータ推定

BT モデルの成り立ちからすると，2 クラス判別器は次の確率を与えることが目的である.

$$q_r(\boldsymbol{x};\boldsymbol{\eta}_r)=p(y\in C_r^+|y\in C_r,\boldsymbol{x}).$$

実際に，$q_r(\boldsymbol{x};\boldsymbol{\eta}_r)$ のパラメータ $\boldsymbol{\eta}_r$ は与えられたサンプルのうち $y\in C_r$ であるものを集め，そのサンプルに対して推定を行なうことになる．したがって $q_r(\boldsymbol{x};\boldsymbol{\eta}_r)$ の訓練用の入力 $\boldsymbol{x}$ は $p(\boldsymbol{x}|y\in C_r)$ にしたがう．一方，異なるクラス，すなわち $y\notin C_r$ となる確率が高い $\boldsymbol{x}$ に関しては $q_r(\boldsymbol{x};\boldsymbol{\eta}_r)$ の返す確率は正確な推論ではないことが予想できる．多クラス判別の際には全てのクラスに対する $\boldsymbol{x}$ の分布 $p(\boldsymbol{x})$ に対して $q_r(\boldsymbol{x};\boldsymbol{\eta}_r)$ の返す値を用いるわけだから，正確な推論でない場合にもその値を用いることになる.

多クラス判別の本来の目的は $p(y|\boldsymbol{x})$ の推論である．モデルは $p_{BT}(y|\boldsymbol{x};\{\boldsymbol{\eta}_r\})$ である．それならば，$\{\boldsymbol{\eta}_r\}$ の推定を $p_{BT}(y|\boldsymbol{x};\{\boldsymbol{\eta}_r\})$ の最尤推定として求める方が自然だ．尤度関数は次の通りである.

$$L(\{\boldsymbol{\eta}_r\})=\sum_t \ln p_{BT}(y_t|\boldsymbol{x}_t;\{\boldsymbol{\eta}_r\}).$$

$L(\{\boldsymbol{\eta}_r\})$ は間接的に $\{\boldsymbol{\eta}_r\}$ の関数となっているが，$p_{BT}(y_t|\boldsymbol{x}_t;\{\boldsymbol{\eta}_r\})$ が陽な関数ではないため，$\boldsymbol{\eta}_r$ に関する最適化は難しそうに見える．以下では，$L(\{\boldsymbol{\eta}_r\})$ の $\boldsymbol{\eta}_r$ による微分をいかに計算するかを示す．微分が求まったならば勾配法によって $\boldsymbol{\eta}_r$ に関する最尤推定を求める.

ここでは 1-vs.-1 判別器すなわち BT モデルを用いることを考える．一般の判別器の組み合せ (GBT モデル) への拡張は難しくない．まず，$\boldsymbol{x}$ を固定して，次の記号を用いることにする.

$$\begin{aligned}&p_k=p_{BT}(k|\boldsymbol{x};\{\boldsymbol{\eta}_r\}),\ \{k\}=C_r^+,\ \{l\}=C_r^-,\\&q_{kl}=q_r(\boldsymbol{x};\boldsymbol{\eta}_r),\ q_{lk}=1-q_{kl}.\end{aligned}$$

$\{p_k\}$ は $\sum_k p_k$=1 の条件の基で (3.3)式で定義された F_{BT} の最大を与えるベクトルであることから，次の関係が成り立つ.

$$\begin{aligned}&\partial_{p_k}\left(F_{BT}+\lambda\left(1-\sum_k p_k\right)\right)=0,\\&\sum_{l\neq k}\left(\frac{q_{kl}}{p_k}-\frac{1}{p_k+p_l}\right)=\lambda,\ k=1,\ldots,K,\end{aligned}\tag{4.1}$$

ここで λ はラグランジュ未定定数である．p_k を両辺に乗じ，k について和をとると

$$\sum_k\sum_{l\neq k}\left(q_{kl}-\frac{p_k}{p_k+p_l}\right)=\lambda=0.$$

したがって $\lambda=0$ となり, (4.1)式から次の K 本の方程式が得られる.

$$\sum_{l\neq k}\frac{p_k}{p_k+p_l}=\sum_{l\neq k}q_{kl},\ k=1,\ldots,K.\tag{4.2}$$

ここで(4.2) 式を η_r で微分する．簡単のため η_r を 1 次元とするが，多次元への拡張は簡単である．もし $\{k,l\}\neq C_r$ ならば右辺は $\partial_{\eta_r} q_{kl}=0$ となる．$\delta_r(k,l)$ は $\{k,l\}=C_r$ ならば 1 を，そうでなければ 0 を返す関数とする．(4.2)式は次のようになる．

$$\sum_{l\neq k}\frac{p_k p_l}{(p_k+p_l)^2}\,(\partial_{\eta_r} l_k-\partial_{\eta_r} l_l)=\delta_r(k,l)\partial_{\eta_r} q_{kl},\quad k=1,\ldots,K. \tag{4.3}$$

ここで $l_k=\ln p_k$ とした．ここでの目的は対数尤度の微分，すなわち $\partial_{\eta_r} l_k$ を求めることである．これは(4.3) 式の線形方程式を解けば得られることになる．

A と $\partial_{\eta_r}\boldsymbol{q}$ とをそれぞれ次のような $K\times K$ 行列，K 次元のベクトルとして定める．

$$A=\{a_{kl}\},\ a_{kl}=\begin{cases}\displaystyle\sum_{l'\neq k}\frac{p_k p_{l'}}{(p_k+p_{l'})^2}, & k=l\\ \displaystyle-\frac{p_k p_l}{(p_k+p_l)^2}, & k\neq l,\end{cases}$$

$$\partial_{\eta_r}\boldsymbol{q}=(\ldots,0,\overset{k}{\underset{\vee}{(\partial_{\eta_r} q_{kl})}},0,\ldots,0,\overset{l}{\underset{\vee}{(\partial_{\eta_r} q_{lk})}},0,\ldots)^T.$$

ここで $\partial_{\eta_r}\boldsymbol{q}$ は k 番目と l 番目の成分のみ値を持ち，他は 0 である．(4.3)式の K 本の方程式は以下のようにまとめることができる．

$$A\begin{pmatrix}\partial_{\eta_r} l_1\\ \vdots\\ \partial_{\eta_r} l_K\end{pmatrix}=\partial_{\eta_r}\boldsymbol{q}.$$

残念ながらこの連立線形方程式は A が正則でないために解を求められない．このことは K 次元の 1 を要素とする行列 $\mathbf{1}=(1,\ldots,1)^T$ を乗じてみればすぐにわかる．

行列 A が縮退する理由は F_{BT} が $\boldsymbol{p}$ の定数倍に関して不変となることによる．そこで，A のうち 1 行を $\sum_k p_k=1$ から導かれる次の方程式で置き換えれば，新たに定義される行列 A' は正則となるため，解を求めることが簡単にできる．

$$\sum_k p_k\partial_{\eta_r} l_k=0.$$

行列 A' とベクトル $\partial_{\eta_r}\boldsymbol{q}'$ は次のように定義される．

$$A'=\{a'_{kl}\},\ a'_{kl}=\begin{cases}a_{kl}, & 1\leq k\leq K-1\\ p_l, & k=K,\end{cases} \tag{4.4}$$

$$(\partial_{\eta_r}\boldsymbol{q}')_k=\begin{cases}(\partial_{\eta_r}\boldsymbol{q})_k, & 1\leq k\leq K-1\\ 0, & k=K,\end{cases} \tag{4.5}$$

連立 1 次方程式は以下のようにまとめられる．

$$A'\begin{pmatrix}\partial_{\eta_r} l_1\\ \vdots\\ \partial_{\eta_r} l_K\end{pmatrix}=\partial_{\eta_r}\boldsymbol{q}'. \tag{4.6}$$

この連立 1 次方程式は A'^{-1} を求めれば簡単に解ける．こうして $\partial_{\eta_r} l_k=\partial_{\eta_r}\log p_{BT}(k|\boldsymbol{x};\{\boldsymbol{\eta}_r\})$ を求めることができる．

最尤推定を求める手順

(1) $q_r(\boldsymbol{x};\boldsymbol{\eta}_r)$ のパラメータ $\{\boldsymbol{\eta}_r\}$ を初期化する．

(2) For $t=1,\ldots,N$,

(a) (3.3) 式を解いて各サンプルに対して以下の確率を求める.

$$\hat{p}_k = p_{BT}(k|\boldsymbol{x}_t;\{\boldsymbol{\eta}_r\}),\ k=1,\ldots,K.$$

(b) $\{\hat{p}_k\}$ を用いて (4.4) 式の A' とその逆行列 A'^{-1} を求める. さらに A'^{-1} の y_t 番目の行を $\boldsymbol{b}^T$ とする.

(c) それぞれのパラメータに対して(4.5)式の $\partial_{\eta_r}\boldsymbol{q}'$ を求める. そして $\boldsymbol{b}$ を用いて次のようにして尤度の微分を求める.

$$\partial_{\eta_r}\ln p(y_t|\boldsymbol{x}_t;\{\boldsymbol{\eta}_r\}) = \boldsymbol{b}\cdot(\partial_{\eta_r}\boldsymbol{q}'). \tag{4.7}$$

(3) (4.7) 式を全ての t について求め, 和をとる. その結果が尤度関数の η_r による微分となる.

$$\partial_{\eta_r}L(\{\boldsymbol{\eta}_r\}) = \sum_t \partial_{\eta_r}\ln p(y_t|\boldsymbol{x}_t;\{\boldsymbol{\eta}_r\}).$$

(4) 単純な勾配法によって η_r を更新する.

$$\eta_r^{new} = \eta_r^{old} + \mu\partial_{\eta_r}L(\{\boldsymbol{\eta}_r^{old}\}), \tag{4.8}$$

ここで μ は小さな正数である.

4.2 数値実験

ここではパラメータ推定を行なうことによる効果を, 簡単な数値実験によって示す. まず, 3 クラスの判別のためのデータを図 1 に示すように生成した. 図中の曲線は最適な判別面を示している. 3 つのうちひとつのクラスのデータは 2 つの領域に分れている. この判別を 1-vs.-1 判別器を用いて BT モデルによって組み合せて行なうことを考える.

まず, 3 つの考えられる 1-vs.-1 判別器をロジスティック回帰によってあらかじめ構成した(図 2, 右). BT モデルによって組み合せたときの判別境界も図に示す(図 2, 左). このときの判別境界は直線となっていることがわかる.

提案した方法によってパラメータを学習した結果を図 3 に示す. それぞれの 1-vs.-1 判別器は学習前とは変化していることがわかる(図 3, 右). これを組み合せた判別境界も学習前とは変化している(図 3, 左). この判別境界についてもパラメータ推定をする前に比べて変化している. 学習以前は直線であった判別境界に対して, 新たに得られた判別境界は直線ではなく曲線となっていることが観察できる.

5. まとめ

2 クラス判別器を組み合せて多クラス判別器を構成する方法は, 実用上効果的な手法であり, 今後の発展が期待されている. 多くの効率的なアルゴリズムが提案され, それらの計算量や数値実験を通じて性能が調べられている. 数値的には有効性が確かめられている一方, それらの間での理論的な比較は十分ではない.

本稿では, これらの手法を統計モデルの推定の問題ととらえ, 比較することが第一の目的であった. その結果, ECOC と BT モデルに基づく手法の間の関係が明らかになった. また, 個々の 2 クラス判別器のパラメータ推定を全体の尤度関数から導き, BT モデルであっても, パラメータの推定が可能であることを示した. 通常, 2 クラス判別器を組み合せる際には各判別器のパラメータは固定されていることが多いが, 本論文の手法を用いれば, パラメータの推定をすることで, 全体の判別の精度を上げられることが示された.

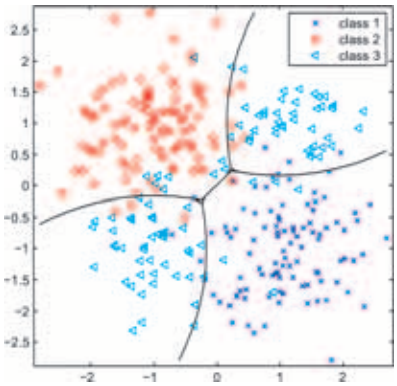

図 1. 3 クラスの判別のためのデータと，教師分布に基づく判別境界.

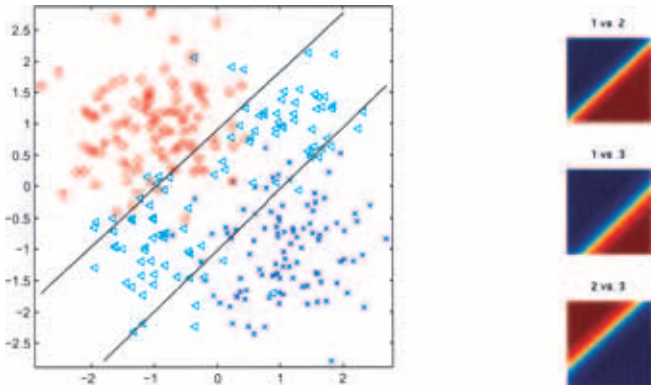

図 2. 個々にパラメータ推定をしたロジスティック回帰モデル (右) とそれらを BT モデルによって組み合せた 3 クラス判別器 (左). 図中の ○, ◁, × で示した点は図 1 に示した 3 クラスの判別のためのデータの再掲である.

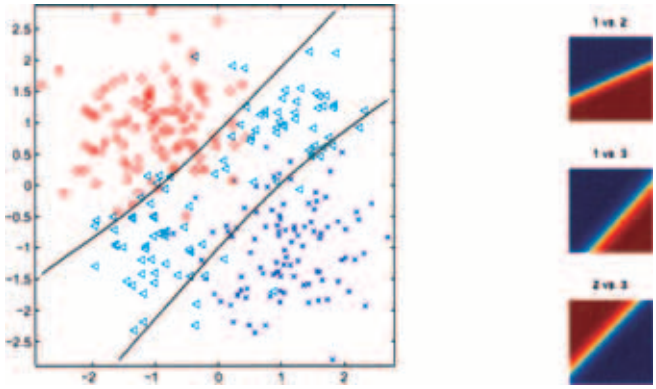

図 3. 提案手法によってパラメータ推定をしたロジスティック回帰モデル (右) とそれらを BT モデルによって組み合せた 3 クラス判別器 (左). 図中の ○, ◁, × で示した点は図 1 に示した 3 クラスの判別のためのデータの再掲である.

謝　辞

論文改訂にあたり査読者から貴重なご意見を頂いた．また，本研究を行なうにあたり 2007 年度まで大学院生として在籍した白石友一君との議論が大変有益であった．ここに記し感謝する．なお，本論文は科学研究費補助金(研究課題番号：18079013)の研究成果に基づくものである．

参 考 文 献

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

Amari, S. (1985). *Differential-geometrical Methods in Statistics*, Lecture Notes in Statistics, Vol. 28, Springer-Verlag, Berlin, Germany.

Boser, B. E., Guyon, I. and Vapnik, V. (1992). A training algorithm for optimal margin classifiers, *COLT*, 144–152.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: The method of paired comparisons, *Biometrika*, **39**, 324–345.

Crammer, K. and Singer, Y. (2002). On the learnability and design of output codes for multiclass problems, *Machine Learning*, **47**, 201–233.

Dietterich, T. G. and Bakiri, G. (1995). Solving multiclass learning problems via error-correcting output codes, *Journal of Artificial Intelligence Research*, **2**, 263–286.

Hastie, T. and Tibshirani, R. (1998). Classification by pairwise coupling, *The Annals of Statistics*, **26**, 451–471.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*, Wiley Series in Probability and Statistics, 2nd ed., Wiley-Interscience Publication, New York.

Huang, T.-K., Weng, R. C. and Lin, C.-J. (2006). Generalized Bradley-Terry models and multi-class probability estimates, *Journal of Machine Learning Research*, **7**, 85–115.

Hunter, D. R. (2004). MM algorithms for generalized Bradley-Terry models, *The Annals of Statistics*, **32**, 384–406.

Ikeda, S. (2006). Learning binary classifiers for multi-class problem, Research Memorandum, No. 1010, The Institute of Statistical Mathematics, Tokyo, http://www.ism.ac.jp/editsec/resmemo-e.html.

Zadrozny, B. (2002). Reducing multiclass to binary by coupling probability estimates, *Advances in Neural Information Processing Systems*, **14**, 1041–1048.

166 Proceedings of the Institute of Statistical Mathematics Vol. 58, No. 2, 157–166 (2010)

Combining Binary Machines for Multi-class: Statistical Model and Parameter Estimation

Shiro Ikeda

The Institute of Statistical Mathematics

Combining binary machines for multi-class classification problems is a popular idea, and many related methods have been proposed. One of the most popular methods is to use the error correcting output codes (ECOC), while another interesting idea is to use the Bradley-Terry (BT) model. In this paper, these methods are reviewed from a statistical model based viewpoint. As a result, a common framework will be given and natural extensions are derived.

Key words: Bradley-Terry model, multi-class classification, maximum likelihood estimation.