4A-03

構成的符号化を用いた ECOC の一構成法（続）

平澤茂一 † 雲居玄道 † 八木秀樹 ‡ 小林学 † 後藤正幸 † 稲積宏誠 §

† 早稲田大学 ‡ 電気通信大学 § 青山学院大学

1 はじめに

本稿では，2 値判別器を複数個（N 個）組合せることで多値（M 値，$M \geq 3$）分類を実現する手法を考える．このような方式は誤り訂正符号の考え方を用いるため ECOC（Error Correcting Output Code）法と呼ばれる．ここでは，低レートで効率の良い Reed-Muller (RM) 符号に注目し，RM 符号を ECOC 法に適するように修正した修正 RM（mRM）符号の生成法を示し，さらに Hadamard 行列を用いて拡張する．これが Simplex 符号と等価な等距離符号であり，これを用いたときの性能を明らかにする．実データや人工データを用い符号長と分類誤り確率を示し，両者のトレードオフ関係をシステム評価の立場から議論する．その結果，カテゴリ数 M が大になるに従い，相対的に効率が良いエラスティック性を持つことを示す．

2 符号語表

2.1 符号語表の構成と性質

ECOC の性能は M 行 N 列の符号語表（Codeword Table）$W = [w_{ij}]$ によって決まる．本稿では $w_{ij} \in \{0,1\}$ の 2 元符号語表を仮定する．W の第 i 行 $\boldsymbol{c}_i = (w_{i1}, w_{i2}, \ldots, w_{iN})$，第 j 列 $\boldsymbol{d}_j = (w_{1j}, w_{2j}, \ldots w_{Mj})^\top$ で表す．ここで，$\top$ はベクトルの転置を示す．

[補元] 任意の長さ L の 2 元ベクトル $\boldsymbol{u} = (u_1, u_2, \ldots, u_L)$ に対し，$u_\ell \oplus u_\ell^{\mathrm{C}} = 1$ を満たす要素 u_ℓ^{C} $(\ell = 1, 2, \ldots, L)$ を持つベクトル $\boldsymbol{u}^{\mathrm{C}} = (u_1^{\mathrm{C}}, u_2^{\mathrm{C}}, \ldots, u_L^{\mathrm{C}})$ をベクトル $\boldsymbol{u}$ の補元 (Complement) と呼ぶことにする．ここで，演算 $\oplus$ は排他的論理和を示す．

2.2 Exhaustive 符号

① 長さ M の 2^m $(m \geq 2)$ 個の列ベクトル $\boldsymbol{d}_j$ に対し，補元 $\boldsymbol{d}_j^{\mathrm{C}}$ を除去する．
② 全 0（または全 1）列ベクトルを除去する．

ここで，①は負例 (0) と正例 (1) を入れ替えれば，同一の判別領域を持つ，②は明らかに判別に寄与していないという意味で冗長である．得られた $N_{\max} = 2^{m-1} - 1$ の $M \times N$ の符号語表は $(N_{\max}, \log_2 M, (M+1)/2)$ Exhaustive 符号を与える．ただし，符号長 N，情報記号数 K，最小設計距離 D の符号を (N, K, D) 符号と示す．

2.3 修正 Reed-Muller (RM) 符号 [1]

任意の正整数 $m (\geq 2)$ に対し，$(2^m, m+1, 2^{m-1})$ 1 次 Reed-Muller (RM) 符号が存在する．ここで，$2M = 2^{(m+1)}$ となる RM 符号を生成し，

① 長さ N の行ベクトル $\boldsymbol{c}_i$ の補元 $\boldsymbol{c}_i^{\mathrm{C}}$ を除去する．
② 全 0（または全 1）を列ベクトル除去する．

得られる修正 RM 符号の M 行 $N (= M-1)$ 列の符号語表は $(M-1, \log_2 M, M/2)$ 修正 RM 符号を与える．ここで，①はどの $\boldsymbol{d}_j$ $(j = 1, 2, \ldots, N)$ においても，$\boldsymbol{c}_i$ と $\boldsymbol{c}_i^{\mathrm{C}}$ は別々のカテゴリとして学習され性能を劣化させ，②は分類に寄与しないという意味で冗長である [1].

3 構成的符号化に基づく符号語表

3.1 修正 RM 符号と Hadamard 行列

本稿の筆者らの 2 人，後藤・小林によって提案された 2.3 で述べた修正 RM 符号は，Plotkin の上界式を等号で満たす数少ない 2 元線形等距離符号である．ただし，$N = 2^m - 1$ でしか存在しない．一方，$\{-1, +1\}$ を要素とする $M \times M$ の Hadamard 行列 H_M に対して $+1$ を 1, -1 を 0 と置き換え，第 1 列の全 0 ベクトルを除去して得られる符号語表は，$N = 2^m - 1$ のとき，やはり 2 元等距離符号を与える．さらに，任意の正整数 ℓ $(\ell \geq 3)$ に対し，$N = 4\ell$ のときに Hadamrd 行列が存在するという仮説があるが，著名な構成方法とともに例が [2][3] に示されている．

3.2 Simplex 符号

$(N, \log_2(N+1), (N+1)/2)$ Simplex 符号は，N 次直交符号から $N-1$ 次に写像した符号として知られ，古くは $(2^m - 1, 2^{m-1} - m, 3)$ Hamming 符号の $(2^m - 1, m, 2^{m-1})$ 双対符号として生成法が知られている．その意味で修正 RM 符号は Simplex 符号の別の生成法を与えている．ただし，$N = 2^m - 1$ である．この Simplex 符号以外にも，$N = 4\ell$ のときに存在している前述した一般の Hadamard 行列による方法を適用すれば，$N = 2^m - 1$ $(m \geq 2)$ および $N = 4\ell$ $(\ell \geq 3)$ の Simplex 符号による ECOC 法の構成が可能である．

3.3 分類性能の解析

構成的符号化または，学習データにより学習した第 j の 2 値判別器 $\boldsymbol{d}_j$ にデータ $\boldsymbol{x}$ を入力する．$\boldsymbol{d}_j$ の出力 $f_j(\boldsymbol{x})$ が真の事後確率を出力する理想的な 2 値判別器の場合，入力 $\boldsymbol{x}$ に対して推定されるカテゴリ $\boldsymbol{c}_{\hat{i}}$ は次式で与えられる [4].

$$\hat{i} = \arg\max_i g(\boldsymbol{c}_i | \boldsymbol{x}). \tag{1}$$

ここで

$$f_j(\boldsymbol{x}) = \sum_{i=1}^{M} w_{ij} \Pr\{\boldsymbol{c}_i | \boldsymbol{x}\} \tag{2}$$

$$g(\boldsymbol{c}_i | \boldsymbol{x}) = \prod_{j=1}^{N} f_j(\boldsymbol{x})^{w_{ij}} (1 - f_j(\boldsymbol{x}))^{1 - w_{ij}}. \tag{3}$$

A Construction Method for ECOC with Constructive Coding

Shigeichi Hirasawa†, Gendo Kumoi†, Hideki Yagi‡, Manabu Kobayashi†, Masayuki Goto†and Hiroshige Inazumi§
†Waseda University
‡The University of Electro-Communications
§Aoyama Gakuin University

4　トレードオフモデルを用いたシステム評価 [5][6]

4.1　トレードオフ関係

一般に ECOC には一定の多値分類数 M（システムの規模）に対し，符号長 N（投資コスト）と分類誤り確率 P（性能の劣化）の 2 律背反（トレードオフ）の関係がある．ただし，( ) 内は一般的システム評価の変数である．ここでは，符号長 N の短縮 Exhaustive 符号の特性を用いて分類問題でしばしば議論されるカテゴリ間の最悪分類誤り確率の下界[*1] P_{ce} を示しながら，Simplex 符号の性能を評価する．$\boldsymbol{x} \in \boldsymbol{c}_i$ が真とすると，下界 P_{ce} は次式で与えられる．

$$P_{ce} = \frac{\sum_{i=1}^{M} P(\boldsymbol{c}_i) \sum_{j \neq i} \Pr\{g(\boldsymbol{c}_j|\boldsymbol{x}) \geq g(\boldsymbol{c}_i|\boldsymbol{x})\}}{M-1}. \quad (4)$$

4.2　システム評価

トレードオフ関係では「わずかな投資コスト増を許容すれば，大きな性能劣化を改善できる」に注目する．これを「フレキシブル」という．ECOC の与えられたシステムの規模 M に対しトレードオフ関係が下に凸な曲線で与えられるとき，

$n = N/N_{\max}$, $p_{ce} = P_{ce}/P_{ce,\max}$ のように正規化し，任意の M に対する比較を可能にする．ここで，$N_{\max}$ ($N_{\min}$) は符号長 N の定義域の最大値（最小値）であり，$N_{\max} = 2^{M-1} - 1$ ($N_{\min} = \lceil \log_2 M \rceil$), $P_{ce,\max} = 1/2$ である．

4.3　人工データ

M 次元多値分類データを平均 $\mu = 1.0$, 分散 $\sigma^2 = 0.1$ の M 次元ガウス分布から生成する．このとき，$M = 4, 8, 12$ の結果を図 1 に示す．ただし，2 値判別器は式 (2) に従う．

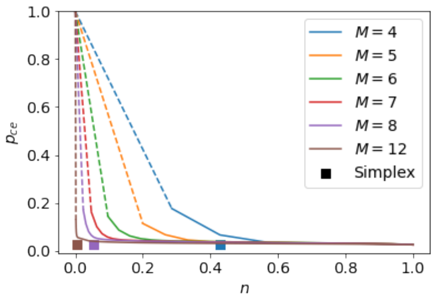

図 1　人工データによるトレードオフ関係

4.4　実データ

手書き数字・英文字 EMNIST データ [7] を用いた結果を図 2 に示す．ただし，2 値判別器は Deep Convolutional Neural Network を用いる．

5　考察

表 1 に Simplex 符号のカテゴリ分類誤り確率 P_{ce} を示す．

1. いずれも Exhaustive 符号に極めて近い P_{ce} を達成する．

次に図 1, 2 より，最悪カテゴリ誤り確率の下界 p_{ce} を示す短縮 Exhaustive 符号の性質を参考に，■で示された Simplex 符号について，以下の結果を得る．

2. Simplex 符号は，M の増加とともに原点方向に向かう「エラスティック」の性質を持つ．
3. Simplex 符号は，M によらずほぼ一定の小さな p_{ce} を持ち，その値は n が小とともに原点方向に向かう．その結果，n は M に対し下に凸な関数を与える．すなわち，「効果的エラスティック」の性質を持つ

[*1] これは特定のカテゴリ間の平均分類誤り確率を与える．

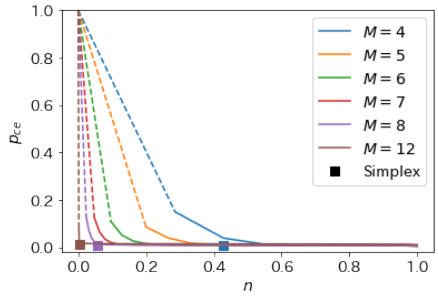

図 2　実データ EMNIST によるトレードオフ関係

表 1　Simplex 符号と Exhaustive 符号による ECOC 法の性能（P_{ce}）

M	人工データ		実データ	
	Simplex	Exhaustive	Simplex	Exhaustive
4	0.0128	0.0128	0.00312	0.00271
8	0.0124	0.0124	0.00259	0.00246
12	0.0124	0.0124	0.00481	0.00506

6　むすび

後藤・小林によって ECOC の性能改善のために提案された修正 RM 符号は線形距離符号であり，$N = 2^m - 1$ の Simplex 符号の生成法の一つであることを示した．Simplex 符号は Hadamard 行列からも生成できることが知られており，そのため符号長が $N \leq 1000$（ただし，668, 716, 892 を除く）の例から多値分類問題の解決には十分実用化可能である．また，Simplex 符号の優れた性質をシステム評価の見方から明らかにし，「エラスティック」「効果的エラスティック」の性質を持つことを示した．なお，Simplex 符号と他の優れた符号を組み合わせた符号語構成法については今後の課題である．

謝辞

本研究の一部は JSPS 科研費 19K04914, 19H01729, 18K11585 の助成による．

参考文献

[1] 後藤正幸, 小林学, 入門パターン認識と機械学習, コロナ社, 2014.
[2] 喜安善市, アダマール行列とその応用, 電気通信学会, 1980.
[3] N.J.A.Sloane, "Library of Hadamard Matrices", http://neilsloane.com/hadamard/.
[4] 雲居玄道, 八木秀樹, 小林学, 後藤正幸, 平澤茂一, "多値分類問題における ECOC 法の最適性に関する一考察", 情報処理学会論文誌 数理モデル化と応用, vol.14, no.3, pp.1—10, 2021.
[5] S.Hirasawa, H.Inazumi, "A system evaluation model by using information theory", *The 30th Joint National Meeting*, MB35.3, 1990.
[6] S.Hirasawa, G.Kumoi, M.Kobayashi, M.Goto and H.Inazumi, "System evaluation of construction methods for multi-class problems using binary classifiers", *The 6th World Conference on Information Systems and Technologies*, 2018.
[7] G.Cohen, S.Afshar, J.Tapson, A.V.Schaik, "EMNIST: Extending MNIST to handwritten letters", *In 2017 International Joint Conference on Neural Networks*, pp. 2921–2926, 2017.