

## 2 値判別器を用いた多値分類方式のシステム評価

### A System Evaluation for Construction Methods of Multiclass Problems using Binary Classifiers

平澤 茂一<sup>(1)</sup>  
Shigeichi HIRASAWA

雲居 玄道<sup>(1)</sup>  
Gendo KUMOI

小林 学<sup>(2)</sup>  
Mababu KOBAYASHI

後藤 正幸<sup>(3)</sup>  
Masayuki GOTO

稲積 宏誠<sup>(4)</sup>  
Hiroshige INAZUM

- (1) 早稲田大学 理工学術院 総合研究所  
Institute of Science and Engineering, Waseda University  
(2) 湘南工科大学 工学部  
Faculty of Engineering, Shonan Institute of Technology  
(3) 早稲田大学 理工学術院 創造理工学部  
Department of Industrial Management Systems Engineering,  
School of Creative Science and Engineering, Waseda University  
(4) 青山学院大学 社会情報学部  
Collage of Informatics, Aoyama Gakuin University

**要旨:**  $N(\geq M-1)$ 個の2値判別器を用いた $M(\geq 3)$ 個のクラス(カテゴリ)からなる多値分類のための構成方法について考察し, レート歪関数の視点より構築したシステム評価モデルによりその性能を評価する. この問題の構成方法は,  $M$ 個(符号語数)の長さ $N$ (符号長)の行ベクトルを持つ $M$ 行 $N$ 列の行列 $W$ をどう与えるかに帰着する. 一般に与えられたカテゴリ数 $M$ に対し, 分類誤り確率 $P_e$ と判別器数 $N$ がトレードオフの関係にあることを示し,  $M$ が大きくなる時の振舞いをシステム評価モデルにより明らかにする. ベンチマークデータ(文書分類問題)により具体的な性能を評価し, 「フレキシブル」「エラスティック」など望ましい性質を持つことを示す.

**Abstract:** Construction methods of the multiple classification systems using binary classifiers are discussed and evaluated by the system evaluation model based on rate-distortion functions. Suppose the multiple classification system constructed by  $M(\geq 3)$  categories and  $N(\geq M-1)$  binary classifiers, then they can be solved by the matrices  $W$ , where the matrices  $W$  are given by the table of  $M$  code words with length  $N$ . Applying the bench-mark data (Newspaper articles of the 2000 Yomiuri Shinbun), the relationships between the probability of classification error  $P_e$  and the number of the binary classifiers  $N$  for a given  $M$  are investigated, and we show that the systems have desirable properties such as "Flexible", "Elastic", and so on.

## 1 はじめに

1970年代後半, J. Pearl と A. Crolotte は質問回答システム(QA システム)の記憶容量と誤りのトレードオフについて, 情報縮約論を用いて解析・評価した[PC79]. すなわち, 各種のQA (Question Answering) システムに対し, 小さな誤りを許容すれば大きな記憶容量を削減できる条件を明らかにした. このような条件を「フレキシブル」(Flexible)や「エラスティック」(Elastic)と呼んでいる. 筆者らはこれを拡張したトレードオフ評価モデルを提案し[Ina89][HI90][HI00], 各種の情報システムに適用した[IKH89][HI00][HSI10]. その結果, 情報システムを設計するに先立って, システムの規模が大きくなるに従い相対的に効率がよい性質「効果的エラスティック」(Effectively Elastic)を持つか否かを判定することが出来る.

一方文書分類や文字認識など, 身の回りには多くの多値分類問題が存在する. 多値分類問題を多値分類器を用いて直接解決する方法もあるが, 一般に計算量が大きく実用的ではない. ここでは, 2値判別器を複数個用いた多値分類方式を考える. また2値判別器として優れた性能を持つSVM (Support Vector Machine) やRVM (Relevance Vector Machine)を導入する. この構成法は誤り訂正符号の構造に着目し, 以前からECOC (Error Correcting Output Codes) として研究されている[DB95][LN10]. ベンチマークデータ(2000年の読売新聞記事)[Yomiuri00]を実データとして用い, 2値分類器を用いた多値分類システム構成部の平均的な性能を評価する. さらに符号理論に立ち戻って, 効率のよい構成方法についても考察する.

本稿では多値カテゴリ数(システムの規模) $M$ をパラメータに, 多値分類システム構成部の判別器数(投資コスト) $N$ に対する分類誤り確率(性能劣化) $P_e$ にトレードオフモデルを適用し, カテゴリ数が増加するに従い望ましい条件を持つか否かを議論する.

## 2 トレードオフ評価モデル

### 2.1 レート歪関数

情報縮約論は有歪の圧縮を対象とし, 情報伝送速度(rate)  $R$  と歪(distortion)  $D$  の関係を論じている. 両者の関係を表すレート歪関数  $R = R(D)$  は一般に下に凸な減少関数で, わずかな歪を許容すれば大きなレートの減少が望める[PC79].

圧縮することにより, 人間の目や耳では認識できない程度の歪を許せば帯域を大きく低減できる. 現在の携帯電話や地上デジタル放送などは, この性質を巧に利用している.

### 2.2 システム評価モデル

J. Pearl と A. Crolotte はレートと歪の間のトレードオフ関係に注目し, QA システムなどの情報システムの評価に適用するモデルを提案した[PC79]. 筆者らはさらにこれを拡張し[HI90][HI00], 大規模システムの設計に際し構成をどのようにすればよいか, あるいは $L$ が増加するとともに, 相対的に効率が良くなる性質を持つかどうかを判定できる基準を提案した. ここで,  $L$  は情報システムの規模を示すパラメータで, レート $R$ はシステムのコストに, 歪 $D$ は性能劣化の程度に対応する.

まず,  $R = R(D)$  を  $D$  軸,  $R$  軸ともそれぞれ  $D_{\max}$ ,  $R_{\max}$  で正規化する. すなわち,  $r = R/R_{\max}$ ,  $d = D/D_{\max}$  とし, 関数

$$r = r(d; L) \quad (2.1)$$

について考える. システムを評価するために, 次の性質を定義する. システム評価関数の式(2.1)について

- (1) 2つのシステムA, Bを考え, それぞれの関数を  $r = r_A(d; L)$ ,  $r = r_B(d; L)$  とする. ある  $L (> 1)$  に対し, 任意の  $d (0 < d < 1)$  において,  $r_A(d; L) > r_B(d; L)$  のとき, システムBはシステムAより「フレキシブル」(Flexible)であるという

- [PC79].
- (2) 任意の  $d(0 < d < 1)$  において、関数  $r = r(d; L)$  の  $r$  が  $L$  の減少関数のとき、「エラスティック」(Elastic) であるという[PC79].
  - (3) (2) において、さらに  $r$  が  $L$  の下に凸な関数のとき、「効果的エラスティック」(Effective Elastic) であるという[HI90].
  - (4) 関数  $d = d(0; L)$  の  $d$  が  $L$  の減少関数のとき、「トリビアルエラスティック」(Trivial Elastic) であるという[PC79]. ただし、 $d = d(r; L)$  は  $r = r(d; L)$  の逆関数である.
  - (5) 任意の  $r(0 < r < 1)$  において、関数  $d = d(r; L)$  の  $d$  が  $L$  の下に凸な減少関数のとき、「マージナルエラスティック」(Marginal Elastic) であるという[HI90].

## 3 2 値判別器による多値分類問題

### 3.1 多値分類システム構成部のモデル

概念学習では、与えられたデータが{yes, no}など 2 値問題を扱う。C4.5 や CART などは木を構成しその葉に 2 値を割り振る。一方手書き文字や新聞記事などの実データの自動分類問題では、分類されるカテゴリは多くが  $M(\geq 3)$  値である。例えば、新聞記事はその掲載される面で{政治, 経済, スポーツ, 社会, ..., 皇室}など 10 値に分類される[Yomiuri00]。例からの学習を考えると、訓練データは  $(\mathbf{x}, C_i)$  の形で与えられる。ここで、 $C_i \in C$  は第  $i$  のカテゴリを示す。(カテゴリが未知の) テストデータ  $\mathbf{y}$  は、訓練データで学習した関数  $f(\cdot)$  を用いて  $f(\mathbf{y}) = C_{i'} \in C$  を求め、 $\mathbf{y}$  が属する第  $i'$  のカテゴリ  $C_{i'}$  を推定する。関数  $f(\cdot)$  は  $M$  値  $\{C_1, C_2, \dots, C_M\} = C$  を取り、 $C$  はカテゴリの集合を示す。表 3.1 に多値分類システムの仕様を示す。なお、軟判定 SVM の第  $j$  の 2 値判別器  $d_j$  は  $h_j(\mathbf{y}) \in (-\infty, +\infty)$  を出力し  $j = 1, 2, \dots, N$  の  $N$  個の判別器を用いて  $f(\mathbf{y})$  を計算する (表 3.1 参照)。2 値分類器を用いた多値分類システム構成部は、ECOC-Matrix [LN10]、判別器構成や符号語構成などと呼ばれる。ここでは符号語表 (Code Word Table) と呼び、次のように行列  $W$  でモデル化する。

$$\begin{aligned}
 W &= [w_{ij}] \quad (w_{ij} \in \{0, 1\}, i = 1, 2, \dots, M, j = 1, 2, \dots, N) \\
 &= [d_1^T, d_2^T, \dots, d_N^T] \\
 &= [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M]^T
 \end{aligned} \tag{3.1}$$

ここで、 $T$  は転置を示す。

表 3.1: 多値分類システムの仕様

|           |                                                                                                                                           |
|-----------|-------------------------------------------------------------------------------------------------------------------------------------------|
| 2 値判別器    | 軟判定 SVM(Support Vector Machine)                                                                                                           |
| SVM の識別関数 | $g(\mathbf{y}) = \omega^T \mathbf{y} + b = \sum_{k=1}^D a_k t_k \mathbf{y} + b$<br>$D$ : 訓練データ数                                           |
| 判別器 (数)   | $d_1, d_2, \dots, d_N$ ( $N$ )                                                                                                            |
| カテゴリ (数)  | $C_1, C_2, \dots, C_M$ ( $M$ )                                                                                                            |
| カテゴリの推定   | $f(\mathbf{y}) = \operatorname{argmax}_{C_j} h_j(\mathbf{y})$                                                                             |
| カテゴリ毎の距離  | $h_i(\mathbf{y}) = \sum_{j=1}^N g(\mathbf{y}) l(w_{ij})$<br>$l(w_{ij}) = \begin{cases} 1 & (w_{ij} = 1) \\ -1 & (w_{ij} = 0) \end{cases}$ |

[例 3.1] 表 3.2a に  $M=5, N=5$ , 表 3.2b に  $M=5, N=7$  の例を示す。前者は one vs. the rest 法と呼ばれ、符号語表の基本型の一つである。

表 3.2a: 符号語表の例 ( $M=5, N=5$ )

|       | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ |
|-------|-------|-------|-------|-------|-------|
| $C_1$ | 1     | 0     | 0     | 0     | 0     |
| $C_2$ | 0     | 1     | 0     | 0     | 0     |
| $C_3$ | 0     | 0     | 1     | 0     | 0     |
| $C_4$ | 0     | 0     | 0     | 1     | 0     |
| $C_5$ | 0     | 0     | 0     | 0     | 1     |

表 3.2b: 符号語表の例 ( $M=5, N=7$ )

|       | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ | $d_6$ | $d_7$ |
|-------|-------|-------|-------|-------|-------|-------|-------|
| $C_1$ | 1     | 1     | 1     | 1     | 1     | 1     | 1     |
| $C_2$ | 1     | 0     | 0     | 1     | 1     | 1     | 0     |
| $C_3$ | 0     | 0     | 1     | 1     | 1     | 0     | 1     |
| $C_4$ | 1     | 0     | 1     | 0     | 0     | 1     | 1     |
| $C_5$ | 1     | 1     | 1     | 0     | 1     | 0     | 0     |
| $d_j$ | 12    | 2     | 8     | 13    | 14    | 11    | 7     |

### 3.2 多値分類問題の実験データ

典型的な多値分類問題として、文書分類を取り上げる。ベンチマークデータとして、2000 年読売新聞記事によるテキスト分類問題[Yomiuri00]を用いる。ベンチマークデータとその実験データの主な仕様を表 3.3 に示す。テキスト分類問題は情報検索の分野で長い歴史があり、実データに基づくベンチマークデータを公開し多くの優れた研究が競い合われて来た。分類問題でも比較的難問とされており、多値分類システム構成部の評価に適していると考えられる。

表 3.3: ベンチマークデータと実験データの仕様

| ベンチマークデータ (2000 年毎日新聞記事) 仕様 |                                                                                                                  |
|-----------------------------|------------------------------------------------------------------------------------------------------------------|
| カテゴリ (数)                    | 10                                                                                                               |
| 文書の形態ベクトル (次元)              | 形態素解析による単語抽出 (3, 059 件)                                                                                          |
| カテゴリとデータ数                   | 政治 26, 053; 経済 20, 642; 社会 12, 744; スポーツ 18, 749; 文化 12, 035; 生活 9, 079; 犯罪事件 23, 787; 科学 3, 184; 国際 235; 皇室 188 |
| 実験データ (抽出し使用したデータ) 仕様       |                                                                                                                  |
| カテゴリ (数)                    | 政治, 経済, スポーツ, 社会, 文化, 生活, 犯罪事件, 科学, (8)                                                                          |
| 実験データ数                      | 合計 4, 400 件                                                                                                      |
| 訓練データ                       | 500 件/カテゴリ, 合計 4, 000 件                                                                                          |
| テストデータ                      | 50 件/カテゴリ, 合計 400 件                                                                                              |

### 3.3 システム評価方法

3.1 でモデル化した対象システムを 2.2 のシステム評価モデルに適用する。多値分類システム構成部に実験データを投入し、2 値判別器数  $N$  を変数としてシステム性能の劣化割合として分類誤り確率  $P_e$  を求める。ここで、問題の複雑さ、すなわちシステムの規模は分類するカテゴリ数  $M$  で与える。情報縮約論・システム評価モデル・対象システムである多値分類システム構成部の変数・パラメータの間に表 3.4 のような対応が成り立っている。

表 3.4: 多値分類システムの評価 (対応表)

| 情報縮約論       | システム評価モデル     | 多値分類システム         |
|-------------|---------------|------------------|
| レート ( $R$ ) | 投資コスト ( $r$ ) | 2 値判別器数 ( $n$ )  |
| 歪 ( $D$ )   | 性能劣化 ( $d$ )  | 分類誤り確率 ( $p_e$ ) |
|             | 規模 ( $L$ )    | カテゴリ数 ( $M$ )    |

## 4 符号語表の作成

### 4.1 Exhaustive 符号の生成

誤り訂正符号は情報記号に検査記号を付加し、通信路やメモリに少々の誤りが生起してもこれを訂正し、信頼度の高い情報記号を再生するために考えられたものである。この考え方を多値分類問題に導入したのが ECOC であり、その最も重要な符号語表が Exhaustive 符号[DB95]である。

この符号語表は全ての  $M$  [ビット] の列ベクトル  $d_j$  ( $j=1, 2, \dots, 2^M$ ) を生成し、その中から

① 補元<sup>1</sup>を取り除く。

② その後、全 1 (または全 0) の列ベクトルを取り除く。

を実行したものである<sup>2</sup>。以上の結果、Exhaustive 符

<sup>1</sup> 元のベクトルの要素の 0 と 1 を入れ替えたもの。

<sup>2</sup> ①は負例(0)と正例(1)を入れ替えれば、結局同じ判別をしていること、②は全 1 (または全 0) の列ベクトルは

号の符号長  $N$ は

$$N = 2^{M-1} - 1 \quad (4.1)$$

である。

[例 4.1]  $M=5$  の Exhaustive 符号の例を表 4.1 に示す。

表 4.1 : Exhaustive 符号 ( $M=5, N_{\max}=15, D=7$ )

|       | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ | $d_6$ | $d_7$ | $d_8$ | $d_9$ | $d_{10}$ | $d_{11}$ | $d_{12}$ | $d_{13}$ | $d_{14}$ | $d_{15}$ |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|----------|----------|----------|----------|----------|----------|
| $C_1$ | 1     | 1     | 1     | 1     | 1     | 1     | 1     | 1     | 1     | 1        | 1        | 1        | 1        | 1        | 1        |
| $C_2$ | 0     | 0     | 0     | 0     | 0     | 0     | 0     | 0     | 0     | 0        | 0        | 0        | 0        | 0        | 0        |
| $C_3$ | 0     | 0     | 0     | 0     | 1     | 1     | 1     | 1     | 0     | 0        | 0        | 0        | 1        | 1        | 1        |
| $C_4$ | 0     | 0     | 1     | 1     | 0     | 0     | 1     | 1     | 0     | 0        | 1        | 1        | 0        | 0        | 1        |
| $C_5$ | 0     | 1     | 0     | 1     | 0     | 1     | 0     | 1     | 0     | 1        | 0        | 1        | 0        | 1        | 0        |

## 4.2 短縮 Exhaustive 符号の構成

$M$  が与えられたとき、符号長を短縮した短縮 Exhaustive 符号<sup>3</sup>を考える。Exhaustive 符号の符号長は式(4.1)で与えられるから、これを  $N_{\max}$  とし  $N_{\min} \leq N \leq N_{\max}$  となる符号長  $N$  の短縮 Exhaustive 符号を考える。ここで  $N_{\min}=M-1$  である<sup>4</sup>。

## 4.3 短縮 Exhaustive 符号の性能評価

Exhaustive 符号はその名の通り、列ベクトルを網羅的 (exhaustive) に抽出している。したがって与えられたシステムの規模 (分類カテゴリ数)  $M$  でコスト (判別器数)  $N$  が最大  $N_{\max}$  である。ここではわずかな性能の劣化 (分類誤り確率  $P_e$  の増加) を許容して、コストを下げる (判別器数  $N$  を減少させる) 短縮 Exhaustive 符号を考える。与えられたシステムの規模  $M$  で、コスト (判別器数、すなわち符号長)  $N$  を変え、その時の性能劣化の割合 (分類誤り確率)  $P_e$  を求める。ここで、得られた曲線は  $N$  に対する  $P_e$  の平均値である。すなわち、符号長  $N$  では  $N_{\max}$  個の列ベクトルから  $N$  個の列ベクトルの組み合わせを全て求めこれを平均している<sup>5</sup>。得られた結果を  $N_{\max}$ ,  $P_{e\max}$  で正規化し、これを  $n = N/N_{\max}$ ,  $p_e = P_e/P_{e\max}$  で示したものが図 4.1 である。

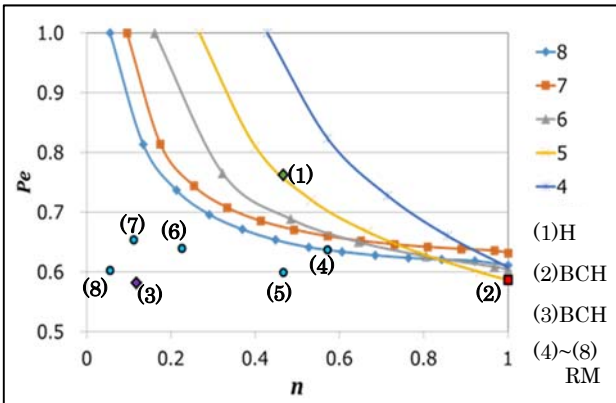


図 4.1 : 投資コスト割合  $n$  と性能の劣化割合  $p_e$  のトレードオフ関係 (パラメータ: システムの規模  $M$ )

判別に寄与していないことから明らかである。

<sup>3</sup> 符号長が  $N$ 、情報記号数が  $K$  の符号を  $(N, K)$  符号と表すとき、符号理論で言う短縮符号は情報記号を  $\alpha$  [ビット] 除去した  $(N - \alpha, K - \alpha)$  符号である。ここでは、情報記号を分離出来る組織符号ではないが、符号長を  $\alpha$  [ビット] 除去して得られる符号を短縮符号と呼ぶことにする。なお以後、最小 (ハミング) 距離  $D$  を特定するとき  $(N, K, D)$  符号で表す。2 つのベクトルのハミング距離とは、異なる要素の数である。

<sup>4</sup> ここでは、比較のため one vs. the rest 型符号語表の列  $(0, 0, \dots, 0, 1)^T$  を除去した修正 one vs. the rest を用いる。

<sup>5</sup> また、符号語とカテゴリの対応をランダムとするために、 $M_0=8$  個から  $M$  個を選ぶ組み合わせを計算し、その平均をとっている。

## 4.4 実験結果と考察

一般に多値分類器では、分類カテゴリ数  $M$  が増加すると分類誤り確率  $P_e$  は増加する。2 値判別器を用いた多値分類システムでも判別器数  $N$  が一定の場合、同様である。ここでは、与えられた  $M$  で  $x, y$  軸をそれぞれの最大値で正規化していることに注意されたい。すなわち、図 4.1 は与えられた  $M$  で、相対的な投資コストと性能の劣化割合のトレードオフ曲線を示している。その結果、両者は

- (1) 下に凸なトレードオフ曲線となり、「フレキシブル」である。
- (2) 一部、 $n \rightarrow 1$  の近傍を除き  $M \rightarrow$  大となるに従いトレードオフ曲線は原点方向に向かい、「エラスティック」である。

## 5 構成的符号語表の作成

4 で議論した短縮 Exhaustive 符号は、ランダムに抽出した列により生成され、その平均的な性能を評価している。図 5.1 に示したように、分類誤り確率を平均化する前には平均値より良い性能を示す構成法 (図中、同じ  $N$  で  $P_e$  が小さいもの) が存在する<sup>6</sup>。このような視点から列ベクトルの選択法に焦点を合わせた研究もある [KKGH16]。

一方 ECOC の分野では、誤り訂正符号を積極的に適用しようとする研究がある [LN10]。その理由は、分類誤り確率を小さくするためには

- (a) 符号語間の最小 (ハミング) 距離を大きくする [DB95] [LN10]。
- (b) (ハミング) 重み<sup>7</sup>が符号語数  $M$  の半分 ( $=M/2$ ) となる列ベクトルを選ぶ<sup>8</sup> [OTKG10]。

などが知られているからである。ここでは 1940 年代からの長い歴史と、多くの理論的・技術的蓄積がある誤り訂正符号の成果を用いて構成的に符号語表を与え、その性能を議論する。誤り訂正符号を用いれば、 $N \rightarrow$  大となる場合テストデータ  $y$  のカテゴリ  $G$  を推定する方法に、優れた代数的復号アルゴリズムを用いて計算量を低減できる可能性がある。

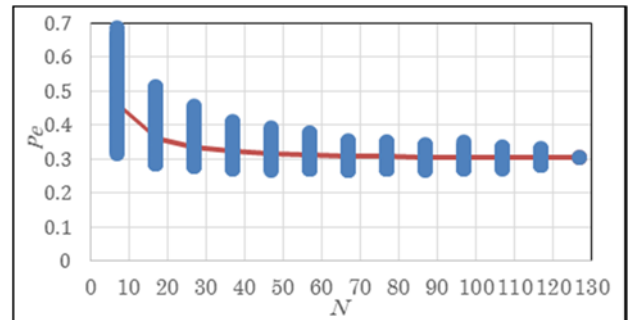


図 5.1: 列ベクトルをランダムに選んだ時の判別器数  $N$  に対する分類誤り確率  $P_e$  の分布 ( $M=8$ )

## 5.1 Hamming 符号と BCH 符号

任意の正整数  $m$  に対し、 $(2^m - 1, 2^m - 1 - m, 3)$  Hamming 符号が存在し、1 [ビット] 誤り訂正符号を与える。同様に、 $(2^m - 1, 2^m - 1 - mt, 2t + 1)$  (原始) BCH 符号が存在し、 $t$  [ビット] 誤り訂正符号を与える。

[例 5.1] Exhaustive 符号との関係

- (1) 実は表 3.2b の  $N=7$ 、 $M=5$  の短縮 Exhaustive 符号は、 $(7, 4, 3)$  Hamming 符号の部分符号<sup>9</sup>で与えられる。表の

<sup>6</sup> 当然、悪いものも存在する。

<sup>7</sup> ベクトルのハミング重みとは、非 0 要素の数である。

<sup>8</sup> 各カテゴリの訓練データ数に偏りがない場合。

<sup>9</sup>  $(N, K, D)$  符号の元の  $2^K$  個の符号語から  $M (\leq 2^K)$  個の符

下部に示した番号は、表 4.1 に示した Exhaustive 符号の列番号に対応している。したがって短縮 Exhaustive 符号の一つである。

- (2) 同様に、 $N=15, M=5$  の Exhaustive 符号は、(15,5,7)BCH 符号の部分符号である。
- (3)  $N=15, M=8$  の短縮 Exhaustive 符号は、(15,5,7)BCH 符号の部分符号である。

ケース(1)(2)(3)の分類誤り確率を求め、図 4.1 に追記している。図中(1)~(3)は上記(1)~(3)のケースに対応している。ただし、ケース(2)は図 4.1 の  $n=1.0$  ( $N=15$ )に一致する。

## 5.2 Reed-Muller (RM) 符号

任意の  $m$  に対し、 $(2^m, m+1, 2^{m-1})$  1 次 Reed-Muller (RM) 符号<sup>10</sup>が存在する。ここで、 $2M=2^{m+1}$  となる RM 符号を生成し、元の符号語（行ベクトル）の補元を除去、全 0（または全 1）の列ベクトルを除去して得られる修正 RM 符号から  $M$  行  $N(=M-1)$  列の符号語表が得られる[GK14]。得られた符号語表から明らかな通り、全ての列ベクトルの重みは  $M/2$  であることが分かる。

[例 5.2] 修正 RM 符号の性能

$M=4$  と、 $M=8$  の修正 RM 符号の符号語表を、それぞれ表 5.1a と b に示す。

表 5.1a : 修正 RM 符号 ( $M=4$ )

|       | $d_1$ | $d_2$ | $d_3$ |
|-------|-------|-------|-------|
| $C_1$ | 0     | 0     | 0     |
| $C_2$ | 1     | 0     | 1     |
| $C_3$ | 1     | 1     | 0     |
| $C_4$ | 0     | 1     | 1     |

表 5.1b : 修正 RM 符号 ( $M=8$ )

|       | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ | $d_6$ | $d_7$ |
|-------|-------|-------|-------|-------|-------|-------|-------|
| $C_1$ | 0     | 0     | 1     | 0     | 1     | 1     | 0     |
| $C_2$ | 1     | 0     | 0     | 0     | 0     | 1     | 1     |
| $C_3$ | 0     | 1     | 0     | 0     | 1     | 0     | 1     |
| $C_4$ | 1     | 1     | 1     | 0     | 0     | 0     | 0     |
| $C_5$ | 0     | 0     | 1     | 1     | 0     | 0     | 1     |
| $C_6$ | 1     | 0     | 0     | 1     | 1     | 0     | 0     |
| $C_7$ | 0     | 1     | 0     | 1     | 0     | 1     | 0     |
| $C_8$ | 1     | 1     | 1     | 1     | 1     | 1     | 1     |

なお、 $M=4,5,6,7,8$  に対し、修正 one vs. the rest と修正 RM 符号の性能を図 4.1 に示している<sup>11</sup>。

## 5.3 実験結果と考察

4 ではランダムに選択した列ベクトルからなる短縮 Exhaustive 符号の平均的な性能を議論している。ここでは特定の構造的符号の部分符号を用いた場合について実験した。結果を[例 5.1]と同様図 4.1 に併記する。図中(4)~(8)は修正 RM 符号の  $M=4\sim 8$  に対応している。また図 5.2 に  $M$  の変化による  $n$  の様子を示す。

- (1) Hamming 符号は、ある  $N$  に対し短縮 Exhaustive 符号の平均的な性能に比べ大差はない。一方、BCH 符号の性能は Hamming 符号に比べ優れている。
- (2) 修正 RM 符号は、平均より、また修正 one vs. the rest の性能より優れている。なお、修正 one vs. the rest の性能は  $p_e=1.0$  のケースより劣る。
- (3) 修正 RM 符号は、 $M \rightarrow$  大とともに  $p_e$  はほぼ一定で  $n$  は減少する。また、ほぼ下に凸の曲線であり「効果的エラストック」である。

号語を選んで得られる符号を部分符号という。

<sup>10</sup> MacDonald 符号とも呼ばれる。

<sup>11</sup>  $M=5,6,7$  については、 $M=8$  の符号語表からランダムに抽出したカテゴリを割り当て、分類誤り確率  $p_e$  を求めその平均を示している。

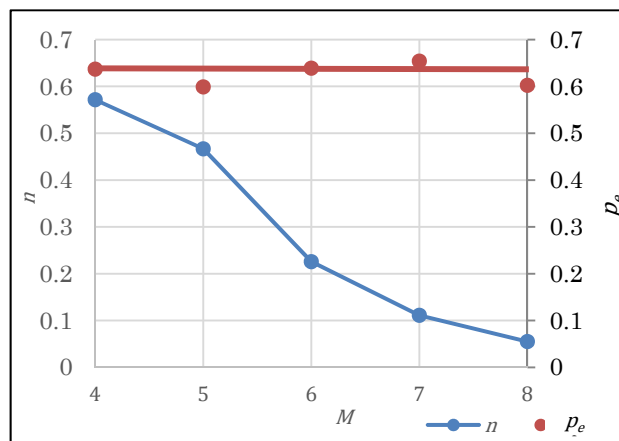


図 5.2 : 修正 RM 符号のシステム規模  $M$  に対する投資コスト割合  $n$  と性能の劣化割合  $p_e$

## 6 むすび

短縮 Exhaustive 符号から出発しその平均的な性能を論じ、次に符号理論の成果を導入して構造的な特定の符号を用いた場合についても議論した。

今後、人工データに対する多値分類システムの詳細な性能評価を行い、その振る舞いを明らかにする予定である。また今回は文書分類問題を用いて評価したが、UCI 機械学習リポジトリのデータセットなど他のベンチマークデータを適用して評価するのも今後の課題である。

**謝辞** 熱心な議論を頂いた早稲田大学・松嶋研・後藤研 D ゼミの皆様へ感謝します。本研究の一部は独立行政法人日本学術振興会学術研究助成基金助成金基盤研究(B) 26282090,(C) 25330045 の助成による。

## 参考文献

- [DB95] T.G. Dietterich and G. Bakiri, "Solving multi-class learning problems via error-correcting out-put codes," Journal of Artificial Intelligence Research, Vol.2, pp.263-286, 1995.
- [GK14] 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 年.
- [HI90] S. Hirasawa, and H. Inazumi, "A system evaluation model by using information theory," 30th Joint National Meeting, ORSA/TIMS, MB35.3, Philadelphia, PE. USA, Oct. 1990.
- [HI00] S. Hirasawa, and H. Inazumi, "A model for syatem evaluation based on information theory," The 2000 International Conference of Management Science and Decision Making, Tamkang University, Taipei, ROC, June 2000.
- [HSI10] 平澤茂一, 斎藤友彦, 稲積宏誠, 「誤り訂正符号を用いた直積ファイルのディスク配置」, 経営情報学会, 2010 年度秋季全国研究発表大会予稿集, B3-1, 名古屋, 2010 年 11 月 6-7 日.
- [IKH89] H. Inazumi, M. Kochiya, and S. Hirasawa, "On the trade-offs between the file redundancy and the communication costs in distributed database systems," IEEE Trans. SMC, vol. SMC-19, no. 1, pp.108-112, Jan. 1989.
- [Ina89] H. Inazumi, Studies on the Evaluations for Information Systems based on Rate Distortion Theory, 早稲田大学博士論文, 1989 年 11 月.
- [KKGH16] 雲居玄道, 小林学, 後藤正幸, 平澤茂一, 「ECOC 法による多値文書分類における符号語構成における一考察」, FIT 2016 第 15 回情報科学技術フォーラム, 富山, 2016 年 9 月 7-9 日. (発表予定).
- [LN10] Y. Luo, and K. Najarian, "Employing decoding of specific error correcting codes as a new classification criterion in multiclass learning problems," Proceedings of 2010 International Conference on Pattern Recognition, pp.4238-4241, 2010.
- [OTKG10] 小田井良輔, 谷口祐樹, 雲居玄道, 後藤正幸, 「事後確率最大判別法に基づく RVM 多値文書分類手法の提案」, 経営情報学会, 2010 年度秋季全国研究発表大会予稿集, C2-1, 名古屋, 2010 年 11 月 6-7 日.
- [PC79] J. Pearl, and A. Crolotte, "Storage space versus validity of answers in probabilistic question answering systems," IEEE Trans. Inform. Theory, vol. IT-26, no. 6, pp.633-640, Nov. 1979.
- [Yomiuri00] 読売新聞記事データ 2000 年版, 日外アソシエーツ株式会社