CA-005

符号理論に基づく多値文書分類における二値判別器の相関に着目した符号語構成法
Construction of Coding Table Method Focused on Correlation of Binary Discriminator in Multilevel Document Classification Based on Coding Theory

雲居玄道* 八木秀樹† 後藤正幸 * 平澤茂一 *
Gendo Kumoi Hideki Yagi Masayuki Goto Shigeichi Hirasawa

1 はじめに

近年，情報化社会の到来により，World Wide Web，電子メール，電子図書館など，膨大なオンラインテキストが扱われるようになった．このような電子媒体のテキストデータを自動処理する技術の重要性は高まる一方であり，中でも高精度な文書自動分類技術が必要とされている．

文書の自動分類技術には様々な手法が提案されているが，特にカーネル法を用いた手法が高性能であると報告されている [1]．その代表的な手法として，Support Vector Machine (SVM) [2] があげられ，優れた二値判別器として知られている．しかし，現実的な問題においては，分類対象となるカテゴリ数が $M(>2)$ となるような多値分類問題が多く，多値分類器の構成法について様々なアプローチから盛んに研究されている．この多値分類器の構成法は大きく分けて 2 つの手法がある．1 つは直接，多値分類を解く問題であり，SVM を多値分類へ拡張した手法 [3] も存在する．もう 1 つの手法として，多値分類問題を二値判別器の集合の構成に落とし込むアプローチが研究されている．これらは，実装コストや計算量などを抑えながら多値分類器を構成することができる．本研究においては，多値文書分類問題に対して，後者の枠組で構成される多値分類器の構成について議論を進める．

その中の方法のひとつとして，符号理論の枠組みを導入した Error-Correcting Output Codes（ECOC）復号法に基づく多値分類法がある [4]．この手法は，各カテゴリを二値判別器の数 $(=N)$ の次元で構成される “符号語” に対応させ，二値判別器の出力結果から符号語を推定しようとするものである．ECOC 復号法に基づく多値分類法を対象とし，効果的な符号構成を議論した研究がいくつかなされている [4], [6], [7], [8]．例えば，符号理論の分野で著名な 2 元符号である BCH 符号や二値判別器の構成をすべて記述した Exhaustive Code を用いた方法 [4]，事後確率の推定誤差を近似的に求めて利用する方法 [6], [7]，多値分類を階層的に構成する方法 [8] などが提案されている．

本研究では，事前に符号語の構成法を与える Exhaustive Code [4] に着目する．この，Exhaustive Code の符号長は，全ての二値判別器の組合せ数であるが，これよりも短い符号長において性能の良い構成が存在する．この点に着目し，その構成法を探索的に発見することを目的とし，ベンチマークデータを用い，その有効性を検証する．

2 通信路と二値判別器

2.1 二元対称通信路

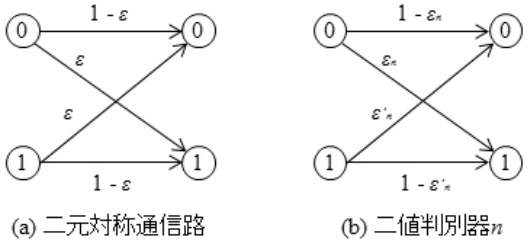
符号理論 [10] において，扱われる通信路モデルの 1 つに二元対称通信路がある (図 1(a))．これは，通信路において送信される記号の集合（アルファベット）が 2 元 $\{0,1\}$ であった場合に，0 (1) を送った場合に誤り確率 ε で雑音が生じ，受信の際に (対称に) 1 (0) と誤ることを意味している．このように，通信路においては，誤りが生じるため受信した系列を送信した符号語に復号する必要があり，本来の情報系列に冗長な情報を付加することにより，それを可能としている．

2.2 二値判別器

Dietterich と Bakiri [4] は符号理論に基づき，多値分類問題を複数の二値判別問題に分解するための枠組みを与えた [4]．N を二値判別器の個数（符号長），M をカテゴリラベル数（符号語数）とした場合，$M \times N$ 行列 $\boldsymbol{W}$ を考える．行列 $\boldsymbol{W}$ の (m,n) 成分を $w_{mn} \in \{0,1\}$ と表す．各行の N 次元ベクトル $\boldsymbol{w}_m$ $(m = 1,2,\cdots,M)$ をカテゴリ C_m の符号語，各列の M 次元ベクトル $\check{\boldsymbol{w}}_n$ $(n = 1,2,\cdots,N)$ を二値判別器 n のカテゴリ分割ベクトルとする．

この時，通信路モデルにおいては，送信されるビットが各二値判別器を通ることから，一般に N 個の二値判別器（ビット位置 n）ごとに誤判別率（図 2．雑音発生確率）が異なる非定常な通信路と捉えることができる．

また，構成される二値判別器において，一般に正例から負例，負例から正例への誤り率，すなわち $0 \to 1, 1 \to 0$ と誤る確率が異なることから，非対称な通信路（図 3）と捉えることができる．このことから，各判別器の誤判別率を $\varepsilon_n, \varepsilon'_n (0 \le \varepsilon_n, \varepsilon'_n < 0.5)$ とした際に，図 1(b) のように表すことができる．



(a) 二元対称通信路　(b) 二値判別器n

図 1．符号理論における通信路と二値判別器

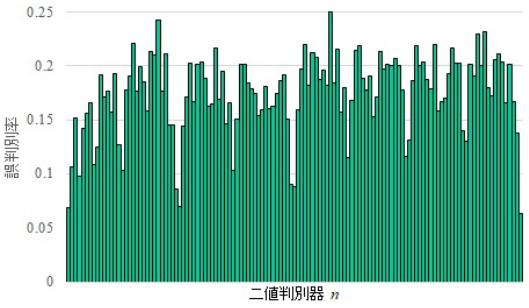


図 2．二値判別器ごとの非定常性 ($M = 8, N = 127$ の例)

*早稲田大学理工学術院
†電気通信大学

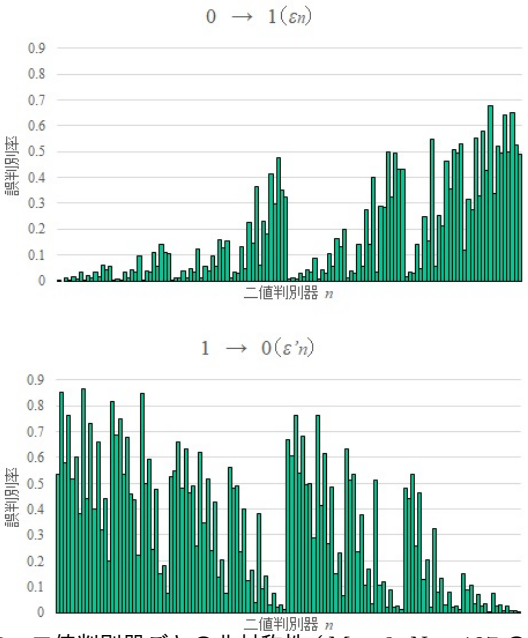


図 3. 二値判別器ごとの非対称性 ($M = 8, N = 127$ の例)

3 復号法

各二値判別器から得られた長さ N の受信系列 $\boldsymbol{r}$ から符号語 $\boldsymbol{w}_m(m = 1, 2, cdots, M)$ への復号においては，M 個の各カテゴリと対応する符号語で構成される $M \times N$ 行列の “符号語表” に対して，入力 $\boldsymbol{q}$ に対する二値判別器の判定結果 $\boldsymbol{r} \in \{0, 1\}^N$ とのハミング距離を計算し，この値が最も近いカテゴリ $\boldsymbol{C}_m$ へ復号するものである [4]．これは，符号理論の分野では “硬判定” と呼ばれる判定方法であり，最も近い符号語が複数あるときに，復号誤りが発生するという問題が発生する．そこで，筆者らは SVM の出力値である分類境界からの距離に着目した復号法を提案した [5]．これは，符号理論の分野では “軟判定” と呼ばれる判別方法の一種であり，一般に硬判定よりも復号誤り率を小さく抑えられる．本研究では，この軟判定手法を前提とし，符号語表の構成方法について議論を行う．

3.1 最小ハミング距離復号（Minimum Distance Decoding）

長さ（符号長）N のベクトル $\boldsymbol{a} = (a_1, a_2, \cdots, a_N)$, $\boldsymbol{b} = (b_1, b_2, \cdots, b_N)$ をその成分毎に比較し，異なる成分の個数を $\boldsymbol{a}$ と $\boldsymbol{b}$ のハミング距離 $D_H(\boldsymbol{a}, \boldsymbol{b})$ という．すなわち

$$D_H(\boldsymbol{a}, \boldsymbol{b}) = \sum_{n=1}^{N} d_H(a_n, b_n) \tag{1}$$

$$d_H(a_n, b_n) = \begin{cases} 0, & a_n = b_n \\ 1, & a_n \neq b_n. \end{cases} \tag{2}$$

である．さらに入力された受信系列 $\boldsymbol{y} \in \{0, 1\}^N$ に対し，符号語数が M の符号語集合を $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_M\}$ とすると，

$$\hat{m} = \arg\min_m D_H(\boldsymbol{x}_m), \boldsymbol{y}) \tag{3}$$

とすると，$\boldsymbol{y}$ を $\boldsymbol{x}_{\hat{m}}$ に復号する方法を最小（ハミング）距離復号法という．

3.2 硬判定復号（Hard Decision Decoding）

判別方法は，符号語 $\boldsymbol{w}_m(m = 1, 2, \cdots, M)$ とテストデータに対する N 個の二値判別器の $\{0, 1\}$ の硬判定出力のハミング距離を H_{C_m} とし，

$$\begin{aligned} \hat{C} &= \arg\min_{C_m} H_{C_m} \\ &= \arg\min_m D_H(\boldsymbol{w}_m, \boldsymbol{r}), \end{aligned} \tag{4}$$

とするカテゴリ $\hat{C}$ に判別する．

この判別方法の場合，ハミング距離が等距離となる符号語が存在した場合には，ランダムに割当てるといった手法がとられる．

3.3 軟判定復号（Soft Decision Decoding）[5]

SVM を用いて硬判定を用いた場合，二値判別器の出力から複数の等距離となる符号語が存在し，誤り検出は可能であるが復号誤りが発生するケースが事前の実験より少なからず（データや判別器性能による）存在することが分かった．このことから，軟判定手法を導入する．SVM を用いた軟判定として，次式 (5) の出力値を距離として用いる軟判定手法が考えられる．

判別方法は，符号語 $\boldsymbol{w}_m$ と入力 $\boldsymbol{q}$ に対する N 個の二値判別器の出力値 $f_n(\boldsymbol{q})$ に対して，

$$G(\boldsymbol{w}_m, \{f_n(\boldsymbol{q})\}_{n=1}^N) = \sum_{n=1}^{N} g(w_{mn}) f_n(\boldsymbol{r}) \tag{5}$$

$$g(a_n) = \begin{cases} 1, & a_n = 1 \\ -1, & a_n = 0. \end{cases} \tag{6}$$

とし，

$$\hat{m} = \arg\max_m G(\boldsymbol{w}_m, \{f_n(\boldsymbol{q})\}), \tag{7}$$

とするカテゴリ $C_{\hat{m}}$ に判別する．

4 符号理論に基づく多値分類法

符号理論において，誤り訂正符号は情報系列にパリティ系列と呼ばれる冗長な情報を付加し，符号語として扱うことにより，情報を伝達する際に多少雑音が混入しても元の情報に訂正することができる符号を指す．

Dietterich と Bakiri は符号理論に基づき，多値分類問題を複数の二値判別問題に分解するための枠組みを与えた [4]．各二値判別器において発生する誤判別を通信路の雑音とみなして，複数の二値判別器を用いることにより誤りを訂正する考え方である．

4.1 Exhaustive Code 構成法

Dietterich と Bakiri が提案した Exhaustive Code [4] の符号語構成方法では，判別器の個数はカテゴリ数から定まり，判別器数は，$N_{MAX} = 2^{M-1} - 1$ 個となる．このとき，各符号語は，

$\boldsymbol{w}_1$ は全て 1 で構成する．$\boldsymbol{w}_2$ は 2^{M-2} 個の 0 に続き $2^{M-2} - 1$ 個の 1 で構成する．$\boldsymbol{w}_3$ は 2^{M-3} 個の 0，2^{M-3} 個の 1，2^{M-3} 個の 0 に続き $2^{M-3} - 1$ 個の 1 で構成する．$\boldsymbol{w}_i$ は 2^{M-i} 個の 0 と 1 を交互に並べて構成する．$M = 5$ の場合の例を表 1 に示す．

表 1. $M = 5$ における Exhaustive Code 符号語表 ($N_{MAX} = 15$)

C_1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C_2	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
C_3	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0
C_4	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0
C_5	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0

この構成法を用いることにより，カテゴリ数 M において，この M 個のカテゴリを 2 つに分割する全ての二値判別器が含まれるように構成することが可能となる一方で，分類器の個数（符号長）が増大するとともに，より短い符号長において性能の良い構成が存在することに着目し，その構成法を発見するすることは必要とされている．

ここで，符号長 N_{MAX} の符号を Exhaustive Code と呼び，$N < N_{MAX}$ となる符号長 N の符号を短縮 Exhaustive Code と呼ぶことにする．

5 提案手法

二値判別器の構成法として，$\{+1, 0, -1\}$ の 3 元で構成し，正例・負例・未使用として構成される手法も存在する [9]．一方で，Exhaustive Code は，$\{0, 1\}$ の二元で構成されており，すべてのデータを使用する形となり，その符号長 $(= N_{MAX})$ は，$2^{(M-1)} - 1$ となることが知られている．本研究では，Exhaustive Code から性能の良い二値判別器の組合せを探索し，多値分類問題に対して有効な短縮 Exhaustive Code による符号語構成法を提案する．

そのため，着眼点のはじめに 2.2 節でも述べたように，二値判別器を組合せた多値分類器を通信路としてみると，非定常・非対称な通信路で構成されているといえる．通信路状態には他にも考慮すべき点があり，それは記憶の有無である．記憶のある通信路とは，送信される各ビットに対して発生する雑音が前のビットと相関があることを意味し，無記憶とは雑音が独立で発生する状態を意味する．これを二値判別器の視点でみると，判別器の構成によっては，二値判別器同士で，ある文書のカテゴリ推定において，誤判別の生起に相関がある状態が考えられる．この点に着目し，二値判別器において発生する誤りの相関に着目した符号語表構成法を提案する．

5.1 複数の判別器への拡張

従来 [11] では，各判別器の非定常性に着目した誤判別率という指標，非対称性に着目した相互情報量という指標を導入した．以下では各判別器に相関があることを考慮して，誤判別率や相互情報量を判別器の組合せに拡張することを考える．このため，誤判別率・相互情報量について，単体の判別器の性能ではなく，複数の判別器の組合せに着目した拡張を行う．

5.2 二値判別器の組合せに拡張した誤判別率

符号語表が $M \times N$ 行列 $\boldsymbol{W}$ として与えられたとき，テストデータが C_m に属する場合，符号語 $\boldsymbol{w}_m$ が送信される符号語となり，二値判別器における推定結果を $\hat{\boldsymbol{w}}_m = \{\hat{w}_1, \hat{w}_2, \cdots, \hat{w}_N\}$ と受信すると見なせる．

このとき，符号語の $n \sim n+l-1$ 番目までの長さ l の部分系列を $\boldsymbol{w}_n^l = (w_n, w_{n+1}, \cdots, w_{n+l-1})$ と表すとき，l 個の連続する二値判別器の組合せにおける送信シンボルと受信シンボルに対応する確率変数を $\tilde{\boldsymbol{w}}_n^l, \hat{\boldsymbol{w}}_n^l$ と書く．この時，受信シンボルにおける組合せを，$\hat{\boldsymbol{s}}^l = (\hat{s}_1, \cdots, \hat{s}_l)$ と表すとき，長さ l の二値判別器の誤判別率は，

$$P_n^l = 1 - \sum_{\hat{\boldsymbol{s}}^l} \Pr\{\tilde{\boldsymbol{w}}_n^l = \hat{\boldsymbol{s}}^l\} \Pr\{\hat{\boldsymbol{w}}_n^l = \hat{\boldsymbol{s}}^l | \tilde{\boldsymbol{w}}_n^l = \hat{\boldsymbol{s}}^l\} \quad (8)$$

と表すことができる．P_n^l は選択された二値判別器の組合せにおける誤判別率である．

5.3 二値判別器の組合せに拡張した相互情報量

相互情報量は，各ビット（判別器）が通信路を介して伝達できる情報量を表す指標である．相互情報量を用いて二値判別器の信頼度を測ることができると考える．シンボルのにおける組合せを，$\boldsymbol{s}^l = (s_1, \cdots, s_l)$ と表すとき，長さ l の二値判別器の相互情報量は，

$$\begin{aligned} I_n^l &= \sum_{\boldsymbol{s}^l} \sum_{\hat{\boldsymbol{s}}^l} \Pr\{\tilde{\boldsymbol{w}}_n^l = \boldsymbol{s}^l\} \Pr\{\hat{\boldsymbol{w}}_n^l = \hat{\boldsymbol{s}}^l | \tilde{\boldsymbol{w}}_n^l = \boldsymbol{s}^l\} \\ &\times \log \frac{\Pr\{\hat{\boldsymbol{w}}_n^l = \hat{\boldsymbol{s}}^l | \tilde{\boldsymbol{w}}_n^l = \boldsymbol{s}^l\}}{\sum_{\boldsymbol{r'}^l} \Pr\{\tilde{\boldsymbol{w}}_n^l = \boldsymbol{r'}^l\} \Pr\{\hat{\boldsymbol{w}}_n^l = \hat{\boldsymbol{s}}^l | \tilde{\boldsymbol{w}}_n^l = \boldsymbol{r'}^l\}} \end{aligned} \quad (9)$$

と定義される．相互情報量は高い方が信頼度が大きい通信路と見なせる．

5.4 組合せの選択方法

N_{MAX} の二値判別器から，式 (8)(9) に基づき，N 個の二値判別器の最適な組合せを求めることは，NP 困難である．そこで，本研究では，貪欲的に選択することを提案する．

貪欲的に選択するとは，事前に選択された二値判別器を元に，誤判別率・相互情報量が最良となるような二値判別器を 1 つ選択することを繰り返す手法である．

いま二値判別器が，$\check{\boldsymbol{w}}_{i_1}, \check{\boldsymbol{w}}_{i_2}, \cdots \check{\boldsymbol{w}}_{i_n}$ 個まで選択された状態を考える．このとき，事前に選択された，$L-1 (< n)$ 個の二値判別器に対して，いまだ選択されていない二値判別器の中から $\check{\boldsymbol{w}}'$ を 1 つを選択し，その状態での式 (8)(9) に基づき誤判別率 P_{n-L-1}^{n+1} または相互情報量 I_{n-L-1}^{n+1} を計算する．

この二値判別器の中から $\check{\boldsymbol{w}}'$ を 1 つを選択することを，選択されていない全ての二値判別器に適用し，最良の二値判別器を次に選択する二値判別器とする．すなわち，$\check{\boldsymbol{w}}_{i_{n+1}} = \check{\boldsymbol{w}}'$ とする．

この時，$L = 1$ のときは，各判別器を独立にみなすことであり，従来 [11] と同様の結果を得る．また，$L = N_{MAX}$ のときには，事前に選択された二値判別器を全て考慮に入れ最良の二値判別器を探索する問題となる．

6 評価実験

本研究では，全てのカテゴリで学習データの数が等しく，データが各カテゴリから出力される確率も全て等しいという問題設定する．

6.1 実験データ

実験データには，2015 年の読売新聞の記事を用いる．

表 1. 実験データ (2015 年読売新聞)

カテゴリ数（数）	政治，経済，スポーツ，社会，文化，生活，犯罪事件，科学 (8)
文書の特徴ベクトル（次元）	形態素解析による単語抽出（60,405 語）
実験データ数	合計 12,000 件
訓練データ	150 件/カテゴリセット，合計 10,800 件
テストデータ	150 件/カテゴリ，合計 1,200 件

表 2 のデータに基づき，実験においては，テストデータとする 1 セットを 10 パターン繰り返して平均をとる 10 分割ローテーションによって評価する．また，式 (8)(9) の値は，それぞれ学習データより推定する．

6.2 比較手法

比較手法として，Exhaustive Code から任意の列となる各分類器を符号長 $N (\leq N_{MAX})$ の数だけランダムに 10,000 回，非復元抽出で選択し構成を決定する手法を用いる．また，直接，多値判別器をモデル化した手法として C-SVC [12] を用いる．

6.3 実験結果

誤判別率に基づく実験結果を図 4，相互情報量に基づく実験結果を図 5 に示す．$L = 1$ とは，各二値判別器の性能を個別にみたものであり，$L = 5$ とは，4 個前の選択された二値判別器に対して，最良となる 1 つを選択していくことである．

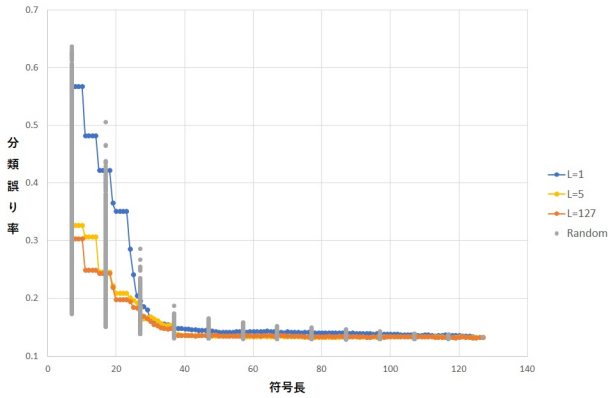


図 4. 誤判別率に基づく分類誤り率

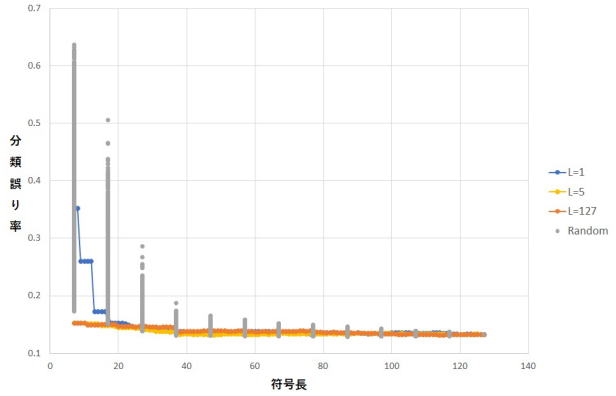


図 5. 相互情報量に基づく分類誤り率

また，各 L において，符号長 $N = 7$〜127 における分類誤り率が最小となるときの符号長および最小値を表 3,4 に示す．また，比較手法として行った C-SVC[12] の平均分類誤り率は，0.500 となった．

表 3. 誤判別率に基づく最小分類誤り率

L	1	5	10	15	20	127
誤り率	0.133	0.133	0.133	0.133	0.133	0.132
符号長	127	123	119	124	123	125

表 4. 相互情報量に基づく最小分類誤り率

L	1	5	10	15	20	127
誤り率	0.133	0.132	0.133	0.133	0.133	0.132
符号長	127	48	120	120	120	115

7 考察

$L = 1$ の各二値判別器の相関を考慮しない選択法においては，図 4,5 より，分類誤り率の低い構成法を符号長が短いところで発見できているが，表 3,4 において，最小誤り率を達成するところが，符号長 127($= N_{MAX}$) となっている．この符号長が 127 となる結果は，Exhaustive Code を用いた時の結果である．

この点，本研究において，提案した相関を考慮した手法を用いることにより，図 4,5 より，相関を考慮しない場合に比べて，符号長が短いところで，良い性能の符号語構成を発見できている．

また表 3 より，Exhaustive Code（符号長 = 127）の誤り率が 0.133 であることから，符号長が短く同等の性能を持つ短縮 Exhaustive Code の符号語構成を発見できている．特に，$L = 5$ のときに，符号長が $N = 48$，分類誤り率 0.132 と短く性能の良い符号語構成を発見することができ，有効な手法であるといえる．

一方で，Exhaustive Code 全ての判別器を一度学習する必要がある点や探索的手法であり最適な符号長が実験的に与えられている点は，改善が必要であると言える．

8 まとめと今後の課題

本研究においては，二値判別器を組合せて多値分類問題に適用する手法として，各二値判別器の性能を通信路状態とみる視点を導入した．この際，各二値判別器は，非定常・非対称である通信路状態であることに加え，相関のある通信路であることを示した．このことに着目し，二値判別器の相関に着目した符号語構成法を提案しその有効性を示した．

今後の課題として，人工データでの実験や Exhaustive Code 全ての判別器を学習せずに，判別器を構成する手法を検討したい．また，Exhaustive Code では，全データを用いた 2 元の符号語表であるが，未使用も含めた 3 元の符号語表への拡張も検討も必要である．

参考文献

[1] B. E. Boser, I. M. Guyon, V. N. Vapnik, “A Training Algorithm for Optimal Margin Classifiers,” *Proceedings of the fifth annual workshop on Computational learning theory*, pp.144–152, 1992.

[2] V. Vapnik and A. Lerner, “Pattern Recognition Using Generalized Portrait Method,” *Automation and Remote Control*, vol.24, pp. 774–780, 1963.

[3] K. Crammer and K. Singer, “On The Algorithmic Implementation of Multiclass Kernel-based Vector-machines,” *Journal of Machine Learning Research*, pp.265–292, Dec. 2001.

[4] T. G. Dietterich and G. Bakiri, “Solving Multiclass Learning Problems Via Error-correcting Output Codes,” *Journal of Artificial Intelligence Research*, vol.2, pp. 263–286, Jan. 1995

[5] 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,” 第 *15* 回情報科学技術フォーラム, F-011, 2016.

[6] 山口暢彦, “WLS-ECOC における事後確率の推定誤差を用いたエラー訂正符号の生成法,” 電子情報通信学会論文誌 *D*, Vol.J89-D, no.2, pp.371–380, 2006.

[7] 白石友一, 福水健次, “多値判別における 2 値判別器のゲーム理論的組合せ法,” 電子情報通信学会論文誌 *D*, Vol.J91-D, no.6, pp.1528–1537, 2008.

[8] 大山賀己, 竹之内高志, 石井信, “ECOC 復号法に基づく階層的多値判別法,” 電子情報通信学会信学技法 *NC2007-542*, Vol.107, pp.337–342, 2008.

[9] E. L. Allwein, R. E. Schapire and Y. Singer, “Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers,” *Journal of machine learning research*, pp.113–141, Dec. 2001.

[10] 平澤茂一，西島利尚, “符号理論入門,” 培風館, 1999.

[11] 雲居玄道，八木秀樹，後藤正幸，平澤茂一, “二値判別器の性能に着目した ECOC 法による多値文書分類における符号語構成に関する一考察,” 情報処理学会第 *79* 回全国大会, 6B-01, 2017.

[12] C. Corinna and V. Vapnik, “Support-vector Networks,” *Machine learning*, pp.273–297, 1995.