情報処理学会論文誌　数理モデル化と応用　Vol.13　No.1　1–12 (Mar. 2020)

多値文書分類のための情報理論的基準による2元符号語表の構成法

雲居 玄道[1,a)]　八木 秀樹[2]　後藤 正幸[1]　平澤 茂一[1]

受付日 2019年1月24日，再受付日 2019年9月25日，
採録日 2019年10月10日

概要：多値分類器の構成法の 1 つに符号理論の枠組を導入した誤り訂正符号に基づく多値分類法がある．本研究では，この枠組に基づく多値分類器の構成に対し，分類誤り率に対して性能の良い符号語構成法を検討する．実データによる実験結果では，冗長な判別器を削除して得られるすべての二値判別器の組合せを並べた Exhaustive Code よりも短い符号長において性能の良い構成が存在する．分類誤り率という点においては，個々の二値判別器の誤り率の非対称性や非定常性などの分類性能だけでなく，二値判別器間の相関も非常に重要な要因になっており，組合せによって全体の分類誤り率は大きく左右される．そこで，二値判別器の相関という視点を導入し，情報理論的基準に基づいた良い組合せの選択法を提案し，符号語の構成法を示す．また，ベンチマークデータを用い，その有効性を検証する．

キーワード：多値分類問題，相互情報量，情報量的基準，2 元符号語表，Error-Correcting Output Codes（ECOC）法，Exhaustive Code，情報理論，二値判別器

Binary Codeword Table for Multilevel Document Classification Using Information Theoretic Criterion

Gendo Kumoi[1,a)]　Hideki Yagi[2]　Masayuki Goto[1]　Shigeichi Hirasawa[1]

Received: January 24, 2019, Revised: September 25, 2019,
Accepted: October 10, 2019

Abstract: A multilevel classification method based on error-correcting codes in a framework of coding theory is one of the powerful construction methods of a multilevel classifier. In this study, we examine the performance for classification error of codeword construction methods for multilevel classifier based on this framework. In fact, there is a configuration with good performance at a code length shorter than the Exhaustive Code in which all combinations of binary discriminators are applied. In terms of classification performance, there are correlations among binary discriminators, and the overall performance depends on the combination of binary discriminators. Therefore, a viewpoint of correlation among binary discriminators is introduced, and search methods based on information theoretic criterion for a good combination of binary discriminators are proposed. We use benchmark data to verify their effectiveness.

Keywords: multilevel classification problem, mutual information, information theoretic criterion, binary codeword table, Error-Correcting Output Codes Method, Exhaustive Code, coding theory, binary discriminators

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

[1] 早稲田大学
Waseda University, Shinjyuku, Tokyo 169–8555, Japan
[2] 電気通信大学
The University of Electro-Communications, Chofu, Tokyo 182–8585, Japan
[a)] moto-aries@ruri.waseda.jp

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（≥ 3）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観点からさか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，多値分類器を構成する方法であり，SVM を多値分類へ拡張した手法 [3] などがある．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研究がある．後者は，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．本研究では，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．

二値判別器を組み合わせて多値分類器を構成する手法において，最もよく知られた方法として，“1-vs-the rest” 法と呼ばれる方法 [4] がある．これは，1 つのカテゴリとそれ以外を識別する二値判別器をカテゴリ数 M だけ用意する方法である．これは簡単な方法である反面，1 つでも二値判別器が誤判別した場合に，所属するカテゴリへ正しく分類できないという問題がある．“1-vs-the rest” 法と同様に，複数の二値判別器の組合せで多値判別を実現する方法として符号理論の枠組を導入した Error-Correcting Output Codes（ECOC）法に基づく多値判別法 [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17] がある．

ECOC 法は，N 個の二値判別器で構成される M 個のカテゴリに対応する “符号語” で構成した “符号語表” を用いて，二値判別器の出力結果全体から符号語を推定するものである．これにより，二値判別器のいくつかで誤判別が生じたとしてもカテゴリを推定することが可能となる．この ECOC 法に基づく多値判別法においては，符号語表の構成法に関する研究 [4], [5], [7], [8], [10], [11], [12], [13], [14] と二値判別器からの出力からカテゴリを推定する復号法に関する研究 [6], [9], [15], [16], [17] がある．

本研究においては，符号語の構成法に着目する．この符号語の構成法においては，事前に用いる符号語表を構成する手法 [4], [5], [6], [7], [8], [9], [11] と構成した符号語表に二値判別器を逐次的に追加していく構成的な手法 [10], [12], [13], [14] がある．このなかでも，符号語表の要素を $\{1, 0\}$ の 2 元として各二値判別器においてすべてのカテゴリを正例または負例に判別する手法 [5], [8] と $\{1, 0, *\}$ の 3 元として，二値判別器によっては正例にも負例にも判別しないカテゴリの存在を許容した 3 元の符号語表を用いる手法 [9], [11] の 2 種類がある．これらの構成法において，明確にデータと性能の関係が明らかになっているとはいい難い．この関係性を明らかにするため，本研究では事前に構成された 2 元の符号語表を用いた ECOC 法 [5] に着目する．この符号語表の構成法として，符号理論の分野で著名な 2 元符号である BCH Code [5], [8] や Reed-Muller Code [4]，すべての二値判別器から構成される含む Exhaustive Code [5]，ランダムに $\{1, 0\}$ を割り当てる Dense Random 法 [6], [7]，1-vs-the rest の法により符号語表を構成する手法 [7], [9] などが提案されている．

ここで，事前に符号語表を与える ECOC 法においては，Exhaustive Code の符号長は等価な二値判別器などの冗長な判別器を削除して得る二値判別器の数（N_{MAX}）に等しい．このことから，任意の符号語表はこの Exhaustive Code の符号語表から二値判別器を削除して得られるものと見なすことができる．ここで M をカテゴリ数とすると，$N_{\mathrm{MAX}} = 2^{M-1} - 1$ と表せる．本研究では，適切な二値判別器を選択し Exhaustive Code より符号長 N（$\leq N_{\mathrm{MAX}}$）が短く分類誤り率が小さい符号語表を構成することを目的とする．従来の符号語表の構成法では，符号語間の距離を指標として，カテゴリに割り当てられる符号語が選択されていた．ここでは各二値判別器の性能が異なることに着目し，情報理論の観点から良い二値判別器の組合せを選択する際に用いる指標を提案する．そして，ベンチマークデータを用いて，提案する指標の有効性を検証する．結果として，提案する指標を用いることにより，少ない二値判別器によって，Exhaustive Code を用いるよりも分類誤り率に対して性能の良い符号語表を得られることを示す．

2. 従来手法

2.1 多値分類問題

分類問題は，新たに与えられた V 次元の入力データ $\boldsymbol{x} \in \mathbb{R}^V$ に対応するカテゴリ $c \in \mathcal{C} = \{C_1, C_2, \ldots, C_M\}$ を推定する問題である．$M \geq 3$ の場合，この問題は，多値分類問題と呼ばれる．このようなカテゴリ推定を行う分類器を実現するにあたり，事前にカテゴリ $l_i \in \mathcal{C}$ の付与された第 i の学習データ $(\boldsymbol{x}_i, l_i)$ の集合を入力データとして学習を行う．文書分類における特徴空間の構成法としては，Bag-of-Words によるベクトル表現がよく用いられる．この方法では，まず学習文書集合内のすべての文書データに対して形態素解析を行い，品詞や頻度による単語選択や不要語の除去の後に形成された単語集合を用意する（単語集合のサイズ $= V$）．次に，各文書に対して，単語集合内の各単語の出現頻度を数え，その頻度情報を要素とするベクトル $\boldsymbol{x}_i$（学習データ）によって文書を表現する．文書ベクトルの要素の計算法としては，単純な頻度以外にも，TF-IDF 尺度のような重み付け法が用いられることもある．

新たに与えられたデータ $\boldsymbol{x}$ のカテゴリの推定においては，まず形態素解析を用いて，文書データを $\boldsymbol{x}$（入力データ）を V 次元の文書ベクトルで表現する．すなわち，この次元 V は，学習データから得られる単語集合サイズで与え

表 1　“1-vs-the rest” 法の符号語表（$M = 5,\ N = 5$）
Table 1　Code word table of “1-vs-the rest” method ($M = 5,\ N = 5$).

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

られる．この文書ベクトルを学習された分類器に入力し，推定されたカテゴリ $C_{\hat{m}}$ を出力として得る．

一般的な多値分類の手法としては，前述のとおり大きく分けて 2 通りの手法が存在するが，本研究では「正例 (1)」と「負例 (0)」の二値に判別する二値判別器を複数組み合わせて多値分類を行う手法を対象とする．文書分類は，単語の種類数は膨大であり，多次元でスパースなベクトルを扱う必要がある問題である．そのため，1 つの分類器によって多値分類を構成する方法を検討するよりも，複数の二値判別器を組合せて処理する方法の方が実装コストや計算量の面で現実的である．また各二値判別器の学習は並列計算が可能になるというメリットがある．

2.2　“1-vs-the rest” 法

二値判別器を組合せて多値分類器を構成する手法の中で最もよく知られた方法として，“1-vs-the rest” 法と呼ばれる方法がある．これは，**表 1** のように，1 つのカテゴリとそれ以外を識別する二値判別器をカテゴリ数だけ用意する方法である．表 1 において，各列は二値判別器に対応し，1, 0 はそれぞれ各二値判別器の正例，負例に対応するラベルを表す．この手法では，未知の文書に対して，各二値判別器から得られる判別結果に基づきカテゴリを推定することができる．たとえば，表 1 の構成に対し，各二値判別器の判別結果が $(1, 0, 0, 0, 0)$ と与えられたならば，C_1 と推定可能である．しかし，これは簡単な方法である反面，1 つでも二値判別器が誤判別した場合，（たとえば $(1, 1, 0, 0, 0)$ のような判別結果が得られた場合）所属するカテゴリを推定できないという問題がある．

2.3　符号理論に基づく多値分類法

2.3.1　Error-Correcting Output Codes（ECOC）法

“1-vs-the rest” 法において，二値判別器の誤判別が生じた場合に所属するカテゴリが推定できない，あるいは誤ったカテゴリに推定されてしまうという問題を解決する手法の 1 つとして，符号理論に基づく多値分類法が提案されている．

符号理論において，誤り訂正符号とは，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を与えた [5]．この枠組に基づく分類手法は ECOC 法と呼ばれる．ECOC 法では，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，二値判別器を冗長に用いることにより誤りを訂正を可能とする．

表 2　$M = 5$ における Exhaustive Code 符号語表（$N_{\mathrm{MAX}} = 15$）
Table 2　Exhaustive Code codeword table at $M = 5$.

C_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	1	1	1	1	1	1	1
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

表 1 において，各カテゴリの系列を符号語としたとき，2 つの符号語間のハミング距離が 2 しかない．したがって，1 つでも誤判別を起こす二値判別器があると，データのカテゴリが推定できないという問題が生じる．そこで，ECOC 法では二値判別器をさらに追加することにより，カテゴリを正しく推定できるようにする．

以降では長さ N の符号語 M 個を行として並べた $M \times N$ 行列 W を符号語表と呼ぶ．行列 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 $\boldsymbol{w}_m$ をカテゴリ C_m の符号語（$m = 1, 2, \cdots, M$），各列の M 次元ベクトル $\boldsymbol{w}^n$ を二値判別器 $n(= 1, 2, \ldots, N)$ のベクトルとする．

2.3.2　Exhaustive Code 構成法

Dietterich と Bakiri は，ECOC 法において Exhaustive Code [5] という符号語表を提案している．この Exhaustive Code は，全カテゴリを二分する二値判別器として考え得るすべてのものから構成されており，符号長（二値判別器数）は，$N_{\mathrm{MAX}} = 2^{M-1} - 1$ となる．$M = 5$ の場合の例を**表 2** に示す．

ここで，任意の各二値判別器 $\boldsymbol{w}^n$（列ベクトル）において，すべての 0 と 1 を入れ替えることは，二値判別器として同じものを表すことになり，一方を削除できることに注意しよう．このため，表 2 の符号語表は，不要な列を削除したすべての二値判別器の組合せから構成されている．

2.3.3　Support Vector Machine

SVM [2] は，分離超平面から最も近いデータまでの距離（マージン）を最大化するように二値判別を行う識別関数の学習手法である．この手法は性能の良い二値判別器として著名な手法であり，性能の良い二値判別器を組み合わせる ECOC 法でよく用いられている．SVM は，マージン最大化によって汎化能力が高いという特徴があり，「高次元特徴空間」，「文書ベクトルの点在性」といった文書分類問題の特性に起因する過学習という問題に対し有効とされている．学習データ $(\boldsymbol{x}_i, l_i)$（$i = 1, 2, \ldots, D,\ \boldsymbol{x}_i \in \mathbb{R}^V,\ V$：特徴ベクトルの次元数）から，各カテゴリ集合の学習デー

タを分離する識別関数 $f(\boldsymbol{x}_i)$ を求める．

入力データ $\boldsymbol{x}$ に対して，二値判別器 $n(=1,2,\ldots,N)$ から識別関数 $f(\boldsymbol{x})$ の出力値として $r_n \in (-\infty,+\infty)$ が出力される．なお硬判定の場合，この出力に対して，$r_n \geq 0$ の場合は正例 $(y_n=1)$，$r_n<0$ の場合は負例 $(y_n=0)$ を出力しカテゴリを判別する．

2.3.4　ECOC 法における復号法

N 個の二値判別器から得られる出力結果を表す長さ N の系列 $\boldsymbol{r}=(r_1,r_2,\ldots,r_N)$ を受信系列と呼び，$\boldsymbol{r}$ からカテゴリを推定する操作を復号と呼ぶ．最も簡潔な復号法は，M 個の各カテゴリと対応する符号語と二値判別器の判定結果 $\boldsymbol{y}=(y_1,y_2,\ldots,y_N)$ とのハミング距離を計算する．そして，この値が最も近いカテゴリ C_m へ復号するものである [5]．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，最小距離復号法である．しかし，最も近い符号語が複数あるときには，復号誤りが発生するという問題が起こる．そこで，筆者らは SVM の出力値である分類境界からの距離（実数値）に着目した復号法を提案した [18]．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，この軟判定復号法 [18] を前提とし，符号語表の構成方法について議論を行う．

本研究で用いる二値判別器である SVM の出力を r_n とする．硬判定復号の場合と異なり r_n は実数値をとる．判別方法は，符号語 $\boldsymbol{w}_m$（行ベクトル）と受信系列 $\boldsymbol{r}$ に対して，

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

$$g(x)=\begin{cases} 1, & x=1 \\ -1, & x=0 \end{cases} \tag{2}$$

とおくとき，

$$\hat{m}=\arg\max_{m} G(\boldsymbol{w}_m,\boldsymbol{r}) \tag{3}$$

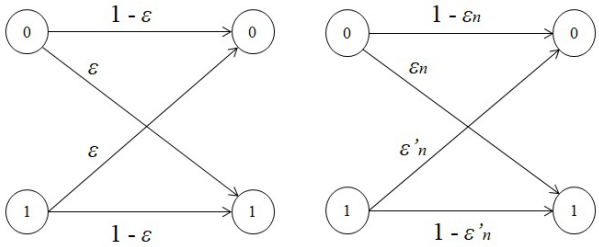
となるカテゴリ $C_{\hat{m}}$ に復号する．

3.　提案手法の着眼点

ECOC 法は，符号理論に基づき二値判別器の誤りを訂正する．本来，符号理論は通信路によって生じる雑音による誤りを訂正するための枠組である．そこで，通信路の雑音と二値判別器による誤りにどのような違いがあるかという点に着目する．

3.1　二元対称通信路

符号理論 [19] 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，0 を送る場合も 1 を送る場合も同じ確率で誤る性質は対称性と呼ばれる．



(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

Fig. 1　Communication channel in coding theory and binary classifier.

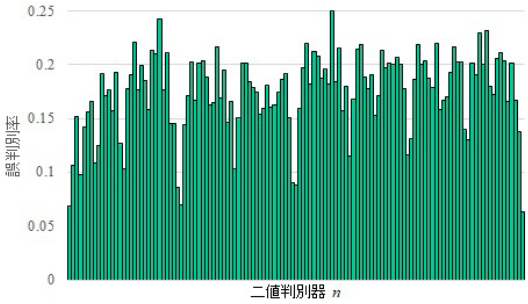


図 2　二値判別器ごとの非定常性（読売新聞，$M=8$，$N=127$ の例）

Fig. 2　Nonstationarity of each binary classifier.

3.2　二値判別器による通信路

3.2.1　非定常性を考慮した基準：平均誤判別率

符号語表が $M\times N$ 行列 W として与えられたとき，カテゴリが既知の D 個の文書は，それぞれカテゴリ C_m に対して，$\boldsymbol{w}_m$ が送信される符号語となり，二値判別器における判定結果を $\boldsymbol{y}=(y_1,y_2,\ldots,y_N)$ と受信すると解釈できる．二値判別器 n 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それぞれ w_n，y_n と書く．このとき，二値判別器 n の平均誤判別率 $\bar{P}_n$ は，

$$\begin{aligned}\bar{P}_n=&\Pr\{w_n=0\}\Pr\{y_n=1|w_n=0\}\\&+\Pr\{w_n=1\}\Pr\{y_n=0|w_n=1\}\end{aligned} \tag{4}$$

と定義される．ここで，$\Pr\{A\}$ は事象 A の確率，$\Pr\{A|B\}$ は事象 B が起きたもとでの事象 A の条件付確率を表す．このとき，$\bar{P}_n$ の値は小さい方が信頼度が高い通信路と見なせる．**図 2** に示すとおり，一般に N 個の二値判別器は判別器（ビット位置 n）ごとに誤判別率（雑音発生確率）が異なる．すなわち，非定常な通信路ととらえることができる（図 1 (b) ε_n，(ε'_n) 参照）．そこで，誤判別率に基づいて二値判別器を選択する符号語表の構成法を提案する．

3.2.2　非対称性を考慮した基準：相互情報量

構成される二値判別器において，一般に負例から正例

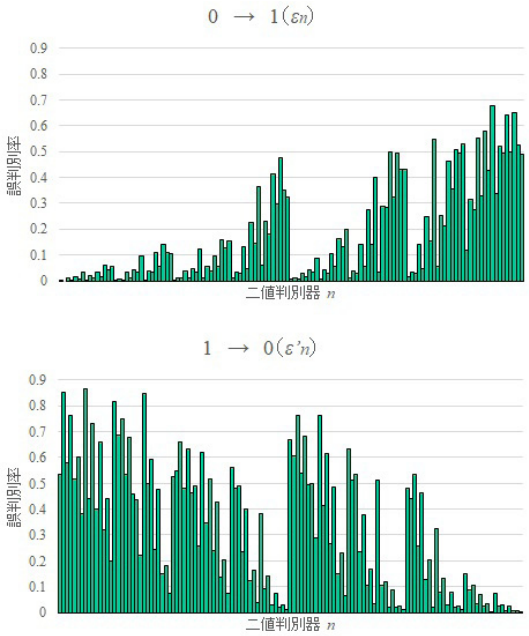


図 3　二値判別器ごとの非対称性（読売新聞，$M = 8$，$N = 127$ の例）

Fig. 3　Asymmetry for each binary classifier.

に誤る事象（$0 \to 1$ と書く）と正例から負例に誤る事象（$1 \to 0$ と書く）の確率が異なることから，各二値判別器を非対称な通信路ととらえることができる（**図 3** 参照）．ここで，$0 \to 1$ と $1 \to 0$ への誤り率をそれぞれ $\varepsilon_n, \varepsilon'_n$ とした際に，図 1 (b) のように表すことができる．

非対称な通信路では，$1 \to 0$ の誤判別率に対して，$0 \to 1$ の誤判別率が大きく（小さく）なることが起こりうる（$\varepsilon_n \neq \varepsilon'_n$）．このことから，二値判別器の誤判別率だけではなく，誤判別の非対称性も考慮した指標の導入が必要であると考えられる．そこで，相互情報量に基づいて二値判別器を選択する符号語表の構成法を提案する．

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．二値判別器を通信路ととらえることで，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．二値判別器 n の相互情報量 I_n は，

$$I_n = \Sigma_{a=0}^{1}\Sigma_{b=0}^{1} \Pr\{w_n = a\} \Pr\{y_n = b | w_n = a\} \times \log \frac{\Pr\{y_n = b | w_n = a\}}{\Sigma_{a'=0}^{1} \Pr\{w_n = a'\} \Pr\{y_n = b | w_n = a'\}} \tag{5}$$

と定義される．I_n は大きい方が信頼度が高い通信路と見なせる．

3.2.3　二値判別器の相関性

通信路には，記憶の有無という性質も存在する．これは，判別器どうしに相関があるか否かということに対応する．

[検証補助実験]

カテゴリ C_m の付与された学習データ $(\boldsymbol{x}_i, l_i)$ が入力されたとき $(l_i = C_m)$，二値判別器 $\boldsymbol{w}^n$ に対して，送信されるビットを w_{mn}，受信するビットを y_{mn} と表す．このとき，式 (6) のように送信ビットと受信ビットの一致を表す変数 z_n を定義する．

$$z_n = \begin{cases} 0 & \text{if } w_{mn} = y_{mn}, \\ 1 & \text{otherwise.} \end{cases} \tag{6}$$

仮に二値判別器の誤りが独立に生起するならば，同時確率と周辺確率の積が一致する．そこで，二値判別器 n_1, n_2 の組に対する同時確率分布と周辺確率分布の積の間の変動距離 $d(n_1, n_2)$ を

$$d(n_1, n_2) = \frac{1}{2} \sum_{a_1, a_2} |\Pr\{(z_{n_1}, z_{n_2}) = (a_1, a_2) | \boldsymbol{w}^{n_1}, \boldsymbol{w}^{n_2}\} - \Pr\{z_{n_1} = a_1 | w^{n_1}\} \Pr\{z_{n_2} = a_2 | \boldsymbol{w}^{n_2}\}| \tag{7}$$

と定義する．変動距離は任意の 2 つの確率分布について 0 以上 1 以下の値をとる距離関数であり，特に 0 となるのは 2 つの確率分布が一致するときに限られる．5.1 節で述べるベンチマークデータを用いてすべての二値判別器の組に対する式 (7) の平均値を算出した．その平均値は 0.502 となった．変動距離の平均値が 0 ではない（特に 0.5 以上の値をとる）ことから，異なる 2 つの二値判別器の誤り方は独立ではない．

そこで，Exhaustive Code から二値判別器を削除して得られる，短縮 Exhaustive Code の構成においては，各二値判別器の性能を個別に計るのではなく，その組合せも考慮する必要があると考えられる．そのため，非定常・非対称性を考慮するとともに，二値判別器の組合せも考慮した選択をする符号語表の構成法を提案する．

4.　提案手法

符号語表の 1 つである Exhaustive Code の符号長は，すべての異なる二値判別器の個数 $2^{M-1} - 1$ に等しい．これに対して，より短い符号長によって多値分類を行うことにより，分類誤り率に対して性能の向上および識別にかかる計算量の削減が可能となると考えられる．そのように作成された符号語表は，Exhaustive Code の部分表現となっている．すなわち，Exhaustive Code から分類誤り率を低下させるように二値判別器を選択することができれば，分類誤り率に対して性能の良い符号語表を構成できると考えられる．ここで，選択された二値判別器によって構成された符号語表を，短縮 Exhaustive Code[*1] と呼ぶ．

本研究では，3 章で提案した 2 つの指標を用いた符号語表の構成法を提案する．

[*1] 符号理論でいう短縮符号は情報記号を α [bit] 除去した符号である．ここでは，情報記号を分離できる組織符号ではないが，符号長を α [bit] 除去して得られる符号を短縮符号と呼ぶことにする．

4.1　二値判別器の組合せに拡張した基準：平均誤判別率

符号語表が $M \times N$ 行列 W として与えられたとき，入力データ $\boldsymbol{x}$ の所属のカテゴリが C_m である場合，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判定結果として $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$ と受信すると見なせる．

符号語 $\boldsymbol{w}$ の n ビット目から $n+l-1$ ビット目までの長さ l の部分系列を $\boldsymbol{w}_n^l = (w_n, w_{n+1}, \ldots, w_{n+l-1})$ と表し，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それぞれ $\boldsymbol{w}_n^l, \boldsymbol{y}_n^l$ と書く．このとき，連続する l 個の二値判別器の誤判別率 P_n^l は，

$$P_n^l = 1 - \sum_{\boldsymbol{a}\in\{0,1\}^l} \Pr\{\boldsymbol{w}_n^l = \boldsymbol{a}\}\Pr\{\boldsymbol{y}_n^l = \boldsymbol{a} | \boldsymbol{w}_n^l = \boldsymbol{a}\} \tag{8}$$

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．

4.2　二値判別器の組合せに拡張した基準：相互情報量

連続する l 個の二値判別器の相互情報量は，

$$I_n^l = \sum_{\boldsymbol{a}\in\{0,1\}^l} \sum_{\boldsymbol{b}\in\{0,1\}^l} \Pr\{\boldsymbol{w}_n^l = \boldsymbol{a}\}\Pr\{\boldsymbol{y}_n^l = \boldsymbol{b} | \boldsymbol{w}_n^l = \boldsymbol{a}\} \times \log \frac{\Pr\{\boldsymbol{y}_n^l = \boldsymbol{b} | \boldsymbol{w}_n^l = \boldsymbol{a}\}}{\sum_{\boldsymbol{a}'} \Pr\{\boldsymbol{w}_n^l = \boldsymbol{a}'\}\Pr\{\boldsymbol{y}_n^l = \boldsymbol{b} | \boldsymbol{w}_n^l = \boldsymbol{a}'\}} \tag{9}$$

と定義される．相互情報量は大きい方が信頼度が大きい通信路と見なせる．

4.3　二値判別器の組合せに拡張した基準：組合せ選択方法

N_{MAX} 個の二値判別器から，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問題は，NP 困難である．そこで，本研究では，事前に選択された二値判別器を元に，誤判別率または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貪欲的な選択手法を提案する．

提案する選択方法では，すでに選ばれている $L-1$ 個の二値判別器 $\{\boldsymbol{w}^{i_{l-L+1}}, \ldots, \boldsymbol{w}^{i_l}\}$ と組合せたときに式 (8)（または式 (9)）を最小（または最大）にする二値判別器 $\boldsymbol{w}^*$ を選択し，$\boldsymbol{w}^{i_{k+1}} = \boldsymbol{w}^*$ とおく．具体的な手順は以下のとおりである．

[選択方法]

N_{MAX} 個のすべての二値判別器の集合を $\mathcal{A}$ とする．いま，すでに二値判別器が l 個まで選択されているとする．

Step 1. すでに選択された二値判別器を $\mathcal{B} = \{\boldsymbol{w}^{i_1}, \ldots, \boldsymbol{w}^{i_l}\}$ とする．

Step 2. $\boldsymbol{w}^* \in \mathcal{A}\backslash\mathcal{B}$ に対し，

$$\mathcal{D}(\boldsymbol{w}^*) = \begin{cases} \{\boldsymbol{w}^{i_1}, \ldots, \boldsymbol{w}^{i_l}, \boldsymbol{w}^*\} & l < L-1, \\ \{\boldsymbol{w}^{i_{l-L+1}}, \ldots, \boldsymbol{w}^{i_l}, \boldsymbol{w}^*\} & l \geq L-1, \end{cases} \tag{10}$$

とおく．式 (8) を最小（または式 (9) を最大）にする $\mathcal{D}(\boldsymbol{w}^*)$ を与える二値判別器 $\boldsymbol{w}^*$ を $\boldsymbol{w}^{i_{l+1}}$ とする．

Step 3. $l = l+1$ とし，Step 1 へ戻る．　■

$L = 1$ のときは，すでに選ばれている二値判別器を考慮せずに $l+1$ 番目の二値判別器を選択することになる．したがって，相関を考慮しない平均誤判別率・相互情報量による構成と同一の結果を得る．また，$L = N_{\mathrm{MAX}}$ のときには，事前に選択された二値判別器をすべて考慮に入れ最良の二値判別器を探索する問題となる．

5.　評価実験

5.1　ベンチマークデータ

ベンチマークデータには，2015 年の読売新聞 [21] および 20 Newsgroups data [22] の記事を用いる．この 2 つのデータは，文書分類問題のベンチマークデータとして広く用いられるものである．特徴としては，読売新聞では，経済，スポーツなどの記事の掲載面レベルでのカテゴリがラベル付けされるのに対し，20 Newsgroups data は，野球やホッケーなどのより細分化されたジャンルがカテゴリラベルとして付与されている．そのため，読売新聞に比べ，20 Newsgroups data は，カテゴリ間の重なりが大きいデータとなっている．どちらも，これらのカテゴリからデータを抽出し，カテゴリ数を $M = 8$ とした．

2015 年の読売新聞は，**表 3** のデータセットに基づき，10 セットから 1 セットを入力データとする実験を 10 パターン行い平均をとる 10 分割ローテーションによって評価する．20 Newsgroups data は，**表 4** のように学習データとテストデータに分割されて与えられている．このデータセットから学習データで分類器を学習しテストデータセットにより評価する．また，式 (8)，(9) の値は，それぞれ学習データより推定する．

L については，各二値判別器を独立に選択する $L = 1$，事前に選択したすべての二値判別器に依存して次が選択さ

表 3　ベンチマークデータ（2015 年読売新聞）[21]
Table 3　Benchmark data (Yomiuri Newspaper, 2015).

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)
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）
実験データ数	合計 12,000 件
データセット	150 件/カテゴリ × 10 セット，合計 12,000 件

表 4　ベンチマークデータ（20 Newsgroups data）[22]
Table 4　Benchmark data (20 Newsgroups data).

カテゴリ（数）	rec.autos，rec.motorcycles，rec.sport.baseball，rec.sport.hockey，sci.crypt，sci.electronics，sci.med，sci.space (8)
文書の特徴ベクトル（次元）	130,107
学習データ	4,762 件
テストデータ	3,169 件

れる $L = 127$ に加えて，$L = 5$ から 5 ずつ変化させた実験を行った．

5.2　比較手法

比較手法として，事前に 2 元の符号語表を構成する Dense Random 法 [6]，BCH Code [5]，Reed-Muller Code [4]，Exhaustive Code [5]，1-vs-the rest [4] を用いる．ここで示した，BCH Code，Reed-Muller Code は，各文献に示されている方法にのっとり構成されるものであり，一般的な符号とは異なっている．また，これらの符号語表も提案手法と同様に短縮 Exhaustive Code に含まれる．以下に各比較手法の詳細を記す．

- Dense Random 法は，列（二値判別器）の重複が生じないように符号語表へ $\{1, 0\}$ をランダムに割当て手法である．これは，Exhaustive Code から列ベクトル（二値判別器）をランダムに非復元抽出で選択し，短縮 Exhaustive Code を構成することと等価である．符号長 N ($\leq N_{\mathrm{MAX}}$) に応じて符号語表を構成することが可能であることから，各符号長ごとに 10,000 回の実験を行い，その平均値で評価する．
- BCH Code を用いた手法は，カテゴリ数よりも多い符号語を生成し，ヒューリスティックに選択する構成法である [5]．本研究においては，(127, 8, 63) BCH Code を用いた．符号語は 2^8 個あるため，ここからランダムにカテゴリ数の符号語を選択し重複する二値判別器を削除した符号語表を短縮 BCH Code と呼びこれを用いる．実験では 10,000 回ランダムに選択し，短縮 BCH Code を生成し，これを符号語表としたときの平均値を計算する．
- Reed-Muller Code を用いた構成法においては，(8, 3, 4) Reed-Muller Code を用いて，0, 1 を反転させると同じになる符号語を削除する．この結果，列の要素がすべて同値になる列が 1 列存在するためこれを削除する．得られる符号語表を**表 5** に示す [4]．

比較手法，提案手法ともに二値判別器は 2.3.3 項で述べた SVM を用い，復号法には 2.3.4 項で記した軟判定復号を用いる．

また，提案手法の貪欲的選択について，最適解との比較を行うため読売新聞のベンチマークデータから 5 カテゴリを抽出し，全数探索[*2]との比較を行う．ここでの全数探索とは，Dense Random 法が取りうる符号語表をランダム構成する手法であることから，この手法を用いて構成可能な符号語表の全数探索を行った．

表 5　Reed-Muller Code を用いた符号語表（$M = 8$）
Table 5　Code word table using Reed-Muller Code ($M = 8$).

C_1	1	1	1	1	1	1	1
C_2	0	0	1	0	1	1	0
C_3	1	0	0	0	0	1	1
C_4	0	1	0	0	1	0	1
C_5	1	1	1	0	0	0	0
C_6	0	0	1	1	0	0	1
C_7	1	0	0	1	1	0	0
C_8	0	1	0	1	0	1	0

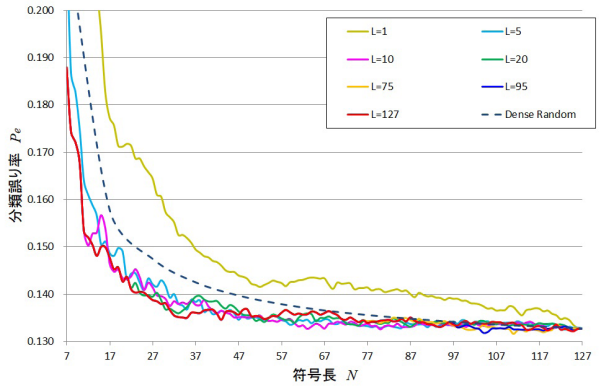


図 4　平均誤判別率に基づく構成法（読売新聞）
Fig. 4　Construction method based on bit error rate (Yomiuri Newspaper).

5.3　実験結果

(1) 平均誤判別率を基準として構成された符号語表を用いた実験結果を**図 4**，**図 6**，**表 6**，**表 8** に示す．
(2) 相互情報量を基準として構成された符号語表を用いた実験結果を**図 5**，**図 7**，**表 7**，**表 9** に示す．
(3) 提案手法における指標と分類誤り率の相関係数を**表 10** に示す．
(4) 各比較手法と提案手法の分類誤り率を**表 11**，**表 12**，

[*2] $M = 5$ では，探索範囲が 2^{15} = 約 3 万であるが，$M = 6$ では，探索範囲が 2^{31} = 約 21 億と爆発的に増大する．このため，5 カテゴリで実験を行った．

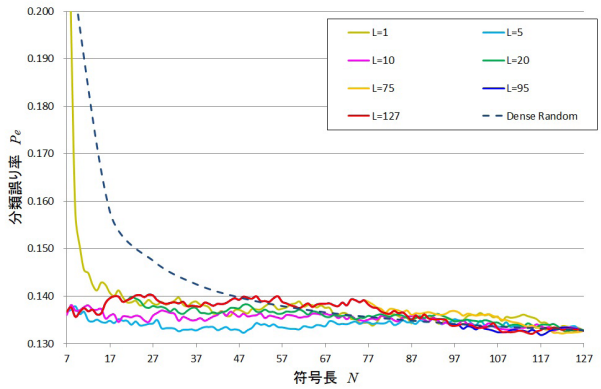


図 5　相互情報量に基づく構成法（読売新聞）

Fig. 5　Construction method based on mutual information (Yomiuri Newspaper).

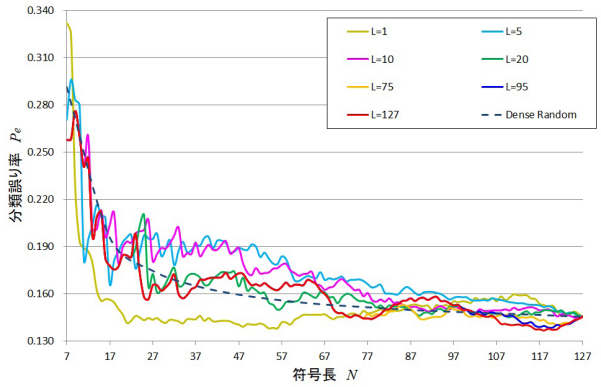


図 6　平均誤判別率に基づく構成法（20 Newsgroups data）

Fig. 6　Construction method based on bit error rate (20 Newsgroups data).

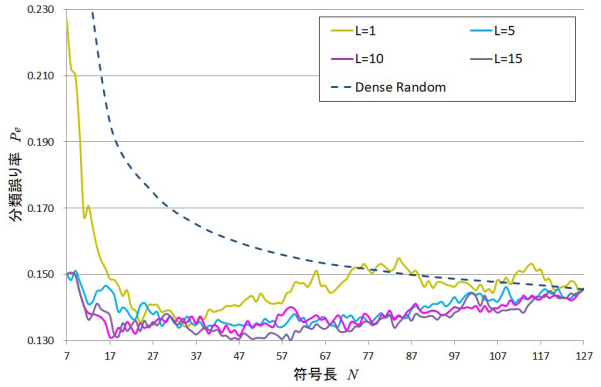


図 7　相互情報量に基づく構成法（20 Newsgroups data）

Fig. 7　Construction method based on mutual information (20 Newsgroups data).

表 6　平均誤判別率に基づく構成法における最小分類誤り率（読売新聞）

Table 6　Minimum classification error rate in construction method based on bit error rate (Yomiuri Newspaper).

L	1	5	10	75	95	127
分類誤り率	0.1328	0.1325	0.1326	**0.1316**	0.1318	0.1322
符号長	127	123	**80**	111	104	125

表 7　相互情報量に基づく構成法における最小分類誤り率（読売新聞）

Table 7　Minimum classification error rate in construction method based on mutual information (Yomiuri Newspaper).

L	1	5	10	75	95	127
分類誤り率	0.1328	0.1323	0.1328	0.1322	**0.1318**	0.1321
符号長	127	**48**	120	122	117	115

表 8　平均誤判別率に基づく構成法における最小分類誤り率（20 Newsgroups data）

Table 8　Minimum classification error rate in construction method based on bit error rate (20 Newsgroups data).

L	1	5	10	50	90	$\geq$115
分類誤り率	0.1381	0.1457	0.1443	0.1398	0.1388	**0.1371**
符号長	**54**	127	126	108	120	118

表 9　相互情報量に基づく構成法における最小分類誤り率（20 Newsgroups data）

Table 9　Minimum classification error rate in construction method based on mutual information (20 Newsgroups data).

L	1	5	10	15	20	$\geq$50
分類誤り率	0.1343	0.1336	0.1312	**0.1302**	0.1315	0.1353
符号長	35	39	**17**	59	56	34

表 10　分類誤り率と平均誤判別率および相互情報量との相関係数（読売新聞）

Table 10　Correlation coefficient between classification error rate and bit error rate and mutual information (Yomiuri Newspaper).

	平均誤判別率	相互情報量
相関係数	0.850	−0.905

表 13，**表 14** に示す．

(5) $M = 5$ における全数探索との比較を**図 8**，**図 9** に示す．

図 4，図 5，図 6，図 7 において，ランダムに選択された短縮 Exhaustive Code の中で，提案手法がどのくらいの分類誤り率を達成するかを評価する．読売新聞を用いた場合は，図 4，図 5 から提案手法においては $N \geq 30$ の範囲で Exhaustive Code（$N = 127$）の分類誤り率に迫る性能を実現していることが分かる．20 Newsgroups data を用いた場合は，図 6，図 7 から提案手法においては Exhaustive Code（$N = 127$）に比べて N が小さい領域で分類誤り率が小さい性能を実現していることが分かる．

なお，符号長 $N = 127$ の点においては，最大の符号長であるためどの手法においても Exhaustive Code となり，その分類誤り率は読売新聞では 0.1328，20 Newsgroups data では，0.1457 となった．

また，分類誤り率が最小となるときの符号長およびその

表 11　比較手法の分類誤り率（読売新聞）
Table 11　Classification error rate of comparison methods (Yomiuri Newspaper).

	比較手法				提案手法	
構成法	Exhaustive Code	BCH Code	Reed-Muller Code	1-vs-the rest	平均誤判別率	相互情報量
分類誤り率	0.1328	0.1344	0.1363	0.1343	**0.1316**	0.1318
符号長	127	78	7	8	111	117

表 13　比較手法の分類誤り率（20 Newsgroups data）
Table 13　Minimum classification error rate of comparison methods (20 Newsgroups data).

	比較手法				提案手法	
構成法	Exhaustive Code	BCH Code	Reed-Muller Code	1-vs-the rest	平均誤判別率	相互情報量
分類誤り率	0.1457	0.1494	0.1560	0.1374	0.1371	**0.1302**
符号長	127	78	7	8	118	59

表 12　比較手法の符号長における提案手法の最小分類誤り率（読売新聞）
Table 12　Minimum lassification error rate of proposed method in code length of comparison methods (Yomiuri Newspaper).

符号長	7	8	78
分類誤り率	**0.1361**	0.1371	**0.1330**
比較手法	RM	1-vs-the rest	BCH

表 14　比較手法の符号長における提案手法の最小分類誤り率（20 Newsgroups data）
Table 14　Classification error rate of proposed method in code length of comparison methods (20 Newsgroups data).

符号長	7	8	78
分類誤り率	**0.1501**	0.1484	**0.1443**
比較手法	RM	1-vs-the rest	BCH

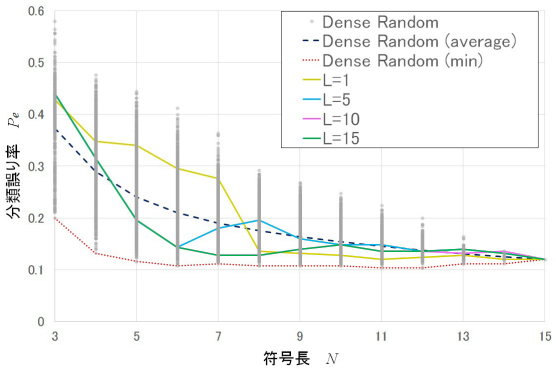


図 8　平均誤判別率に基づく構成法（読売新聞，$M = 5$）
Fig. 8　Construction method based on bit error rate (Yomiuri Newspaper, $M = 5$).

最小値を表 6，表 7，表 8，表 9 に示す．最小の分類誤り率のときの符号長が $N = 127$ ということは，短縮 Exhaustive Code が Exhaustive Code と同等かそれ以下の性能であることを示している．

最後に，表 11，表 13 に比較手法の結果と提案手法の結果を示す．符号長が任意に決定可能な提案手法は，平均誤判別率，相互情報量それぞれで，最小の分類誤り率となるものを示す．合わせて，比較手法と同じ符号長とした場合の提案手法の分類誤り率を表 12，表 14 に示した．

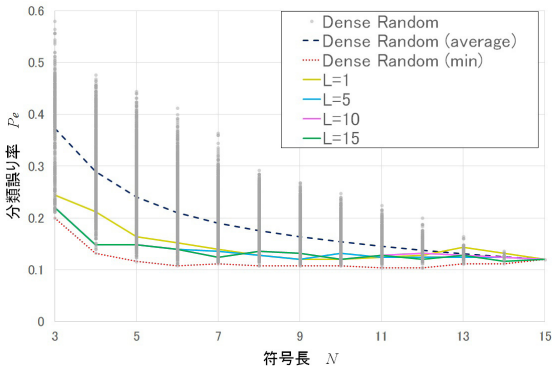


図 9　相互情報量に基づく構成法（読売新聞，$M = 5$）
Fig. 9　Construction method based on mutual information (Yomiuri Newspaper, $M = 5$).

5.4　結果の解釈と考察

結果の解釈と考察について，以下に評価に用いたベンチマークデータごとに述べる．

5.4.1　読売新聞を用いた場合

(a) 平均誤判別率を用いた提案手法

表 11 より，Exhaustive Code（$N = 127$）の分類誤り率が 0.1328 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 が存在することが分かる．提案手法での最も性能が良い符号語表は，$L = 75$，$N = 111$ のときに分類誤り率が 0.1316 で最小となる．

(b) 相互情報量を用いた手法

表 11 より，提案手法での最も性能が良い符号語表は，$L = 95$，$N = 117$ のときに，分類誤り率が 0.1318 で最小となる．

(a) と (b) を比較すると，両者は符号長，分類誤り率の点からほぼ同等の性能と考えられる．ただし，計算量では，

誤りの種類を区別しない (a) 平均誤判別率を用いた方が少ない．

5.4.2　20 Newsgroups data を用いた場合

(a) 平均誤判別率を用いた提案手法

表 13 より，Exhaustive Code（$N = 127$）の分類誤り率が 0.1457 であることから，符号長が短く性能の良い短縮 Exhaustive Code が存在することが分かる．提案手法での最も性能が良い符号語表は，平均誤判別率に基づく構成法では $L = 115$，$N = 118$ のときに分類誤り率が 0.1371 で最小となる．その結果，入力された未知文書のカテゴリ判定において，$N_{MAX} = 127$ 個の二値判別器を使用する Exhaustive Code よりも性能の向上が見込まれる．

(b) 相互情報量を用いた手法

表 13 より，提案手法での最も性能が良い符号語表は，$L = 15$，$N = 59$ のときに，分類誤り率が 0.1302 で最小となる．Exhaustive Code（$N = 127$）の分類誤り率が 0.1457 であることから，大きな性能向上がみられた．

5.4.3　データセットに対する性能の比較

Dense Random 法を用いて特定の符号長ごとにランダムに 10,000 回ずつ探索を行った際に，個々の探索された短縮 Exhaustive Code を比較してみると，読売新聞では 8.9%が，20 Newsgroups data では 15.2%が，Exhaustive Code より良い性能を示した．このことから，20 Newsgroups data は，相互情報量を用いた提案手法による短縮 Exhaustive Code により良い性能が得られやすいことが分かる．これはベンチマークデータの分類問題としての難易度に関わりが大きいと考えられる．読売新聞では，スポーツと政治などを分ける分類問題であるのに対し，20 Newsgroups data は，野球やホッケーといった，同じスポーツに含まれるカテゴリがあるため，カテゴリ間の重なりが大きい（類似性の高い）データセットと考えられる．このため，分類問題として難しい課題となっている．これは，同一の符号語表を用いた Exhaustive Code（$N = N_{\mathrm{MAX}}$，$M = 8$）の結果の分類誤り率は読売新聞が 0.1328，20 Newsgroups data は 0.1457 である．すなわち，分類問題としての難易度は，20 Newsgroups data の方が高いことが実際に分かる．このような問題に対して，短縮 Exhaustive Code を用いて多値分類器を構成することで，性能が向上することが期待できる．

5.4.4　提案手法の 2 つの基準の比較

カテゴリ数 $M = 8$ のケースにおいて，最小の分類誤り率を達成している L は，$L > 1$ であることから，相関を考慮しない $L = 1$ に比べて，提案手法の相関を考慮する指標の有効性が確認できる．最小の分類誤り率で比較すると，読売新聞では平均誤判別に基づく構成法であるが，20 Newsgroups data では，相互情報量に基づく構成法であった．これは，20 Newsgroups data が複雑な分類問題になっているためと考えられる．平均誤判別率の計算量は相互情報量よりも少ない．そのため，実用上メリットである．

表 15　最小分類誤り率（読売新聞，$M = 5$）

Table 15　Minimum classification error rate (Yomiuri Newspaper, $M = 5$).

構成法	Dense Random（min）（全数探索）	提案手法（平均誤判別率）（$L = 1$）	提案手法（相互情報量）（$L = 15$）
分類誤り率	**0.104**	0.120	0.116
符号長	11, 12	11, 14	14

一方で，図 5，図 7 に着目すると，データセットによらず，図 4，図 6 に対して符号長 N が小さいところで，急速に分類誤り率が小さくなっている．この結果は，非定常性に着目する平均誤判別よりも，より詳細に各二値判別器の性能を評価する非対称性まで考慮した，相互情報量が重要であることを示している．

5.4.5　提案指標の妥当性

提案した指標が分類誤り率を改善するものとなっているか確認するため，“Random” で探索した符号語表に対して，提案した式 (8)，(9) に基づき算出した値と分類誤り率の相関係数を表 10 に示した．これにより，提案した指標が分類誤り率に対して強い相関があり，有効な指標であるといえる．

一方で，本研究で提案した指標に基づいて探索する手法には，$M = 5$ で貪欲的な選択を用いた．このため，最適解の保証はない．そこで，貪欲的選択と全数探索の比較を図 8，図 9 および**表 15** に示した．ここで Dense Random 法を用いて全数探索を行ったことから，図中の Dense Random (min) が達成可能な最小分類誤り率となる．$L = 1$ は組合せを考慮しない手法であり，すでに選択された二値判別器に依存して選択される提案手法（$L > 1$）の貪欲的選択において，最適な最小分類誤り率に近い結果を得られていることが分かる．このことより，最適解を求めるために膨大な計算量がかかる $M = 8$ においても，提案手法の貪欲的な選択が有効であると考えられる．一方で，表 15 より最適解に近い組合せを探索する余地が残されていることも分かる．

5.4.6　比較手法との同一符号長での性能

表 11，表 13 より，すべての場合で提案手法は分類誤り率の点から比較手法より優れている．さらに，比較手法に対して同じ符号長での提案手法の性能は，表 12，表 14 のようになる．この結果から，同一の符号長でも BCH Code（$N = 78$）および Reed-Muller Code（$N = 7$）の構成法より性能が良いことが分かる．ただし，1-vs-the rest 法（$N = 8$）と比べると提案手法の性能が悪くなっている．また，符号語表を学習することなく構成的に与える分類問題に対して，固定長で短い符号長での性能を追求した場合には，1-vs-the rest 法や Reed-Muller Code に注目すべきであることが示唆された．ただし，Reed-Muller Code はカ

テゴリ数 M の値の選択に自由度が小さいことに注意する必要がある [4].

5.4.7　提案手法の課題

平均誤判別率および相互情報量の算出には，Exhaustive Code すべての判別器を 1 度学習させる必要がある．また，最適な符号長が実験的にしか与えられない点は，改善が必要であるといえる．

6. 考察

本研究では，多値文書分類問題に対して，二値判別器の組合せにより分類器を構成する手法に対して提案を行い，文書データに対して性能評価を行った．その結果として，最小分類誤り率は 0.1302 であった．この値は，専門家により人手で分類すれば，より小さくできると考えられる．医療分野など対象とする文書によっては，分類誤り率を 1 割以下にすることがが要求されることから，よりいっそうの改善が望まれる．

提案手法は，20 Newsgroups data のようなカテゴリ間の類似性の高いデータセットでより有効であることが分かった．同様の性質を持つ分類問題として，文書データ以外では，代表的なものに手書き文字認識がある．この分類問題においては，「あ」と「お」などの似ている文字はまとめて分類したうえで，個別に分類問題を構築する階層的な手法が取られている．このように，誤分類の発生しやすいカテゴリが多数の組合せで存在する問題においては，それらを同一のカテゴリと見なすことで，二値判別器の性能が向上することから，本手法はより性能を発揮すると予想される．

一方で，複数のカテゴリにまたがる包括的なカテゴリのある分類問題やデータ数に大幅な偏りのある分類問題も考えられる．このような分類問題に対し，本研究で用いたデータ全体の分類誤り率を基準とすると，平均的に性能が良い符号語表であったとしても，カテゴリ間で性能のばらつきが大きいことがありうる．そのため，カテゴリの構成によっては，本研究で用いた全カテゴリでの分類誤り率を基準としたマクロ平均ではなく，カテゴリごとの分類誤り率を基準とするミクロ平均などを用いることも必要である．

そして，本研究で提案した指標では，すべての二値判別器の学習が必要となる．本研究で取り扱ったカテゴリ数 8 以下であれば学習時間は微小である．また，二値判別器は並列に学習可能であることから一定のカテゴリ数までは対応可能であると考えられる．しかし，カテゴリ数が増えると指数的にこの学習時間は増大する．たとえば，カテゴリ数が 1,000 などと非常に大きな分類問題においては，すべての二値判別器を学習することは現実的ではなく，指標の見直しが必要である．その上，ECOC 法では，データの分類の際に一般に各カテゴリの符号語と比較する必要がある．このため，カテゴリ数が増大することにより，カテゴリの推定時間も増加する．これに対応するため，符号理論で用いられる代数的復号が可能な符号語表の導入が考えられる．

7. まとめと今後の課題

本研究では，二値判別器を組み合せて多値文書分類問題に適用する手法に対し，各二値判別器を通信路とみる視点を導入した．この際，分類対象となるデータがすべての二値判別器に同様に入力されるため，各二値判別器は非定常・非対称な通信路であることに加え，互いに相関のある通信路であることを示した．これらの性質に着目し，本研究では情報理論の観点から良い二値判別器の組合せを選択する際に用いる指標を提案し，その有効性を示した．特に，それぞれの指標で二値判別器の相関を考慮すれば，Exhaustive Code と同等の分類性能を持つ短い符号長の符号語表が構成できることを示した．

今後の課題として，これらの指標をもとに，符号理論の知見と合わせて符号語表を構築する方法などがあげられる．加えて，画像認識などの文書以外のデータでの提案手法の検証も必要である．

最後に，本研究での対象は 2 元符号語表とした．一方，カテゴリによっては正例にも負例にも判定しないことを許容する 3 元符号語表を用いる ECOC 法も応用上は重要である．たとえば [6], [9], [12] の多値分類手法は 3 元符号語表を用いる ECOC 法と見なすことができる．提案した二値判別器の選択基準は 3 元符号語表を用いる場合にも適用できるが，その際の性能の評価についても今後の課題である．

参考文献

[1] Boser, B.E., Guyon, I.M. and Vapnik, V.N.: A Training Algorithm for Optimal Margin Classifiers, *Proc. 5th Annual Workshop on Computational Learning Theory*, pp.144–152 (1992).

[2] Vapnik, V. and Lerner, A.: Pattern Recognition Using Generalized Portrait Method, *Automation and Remote Control*, Vol.24, pp.774–780 (1963).

[3] Crammer, K. and Singer, K.: On The Algorithmic Implementation of Multiclass Kernel-based Vectormachines, *Journal of Machine Learning Research*, Vol.2, pp.265–292 (2001).

[4] 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).

[5] Dietterich, T.G. and Bakiri, G.: Solving Multiclass Learning Problems Via Error-Correcting Output Codes, *Journal of Artificial Intelligence Research*, Vol.2, pp.263–286 (1995).

[6] Allwein, E.L., Schapire, R.E. and Singer, Y.: Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers, *Journal of Machine Learning Research*, Vol.1, pp.113–141 (2001).

[7] Escalera, S., Pujol, O. and Radeva, P.: Error-Correcting Output Codes Library, *Journal of Machine Learning Research*, Vol.11, pp.661–664 (2010).

[8] Windeatt, T. and Ghaderi, R.: Coding and Decoding

Strategies for Multi-Class Learning Problems, *Information Fusion*, Vol.4, No.1, pp.11–21 (2003).
[9] 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.91, No.6, pp.1528–1537 (2008).
[10] Rocha, A. and Goldenstein, S.K.: Discriminant ECOC: Multiclass from binary: Expanding one-versus-all, one-versus-one and ecoc-based approaches, *IEEE Trans. Neural Networks and Learning Systems*, Vol.25, No.2, pp.289–302 (2006).
[11] Pedrajas, N. and Boyer, D.: Improving Multi-class Pattern Recognition by the Combination of Two Strategies, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol.28, No.6, pp.1001–1006 (2006).
[12] Pujol, O., Radeva, P. and Vitria, J.: Discriminant ECOC: A Heuristic Method for Application Dependent Design of ECOC, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol.28, No.6, pp.1007–1012 (2006).
[13] Escalera, S., Pujol, O. and Radeva, P.: Boosted Landmarks of Contextual Descriptors and Forest-ECOC: A Novel Framework to Detect and Classify Objects in Cluttered Scenes, *Pattern Recognition Letters*, Vol.28, No.13, pp.1759–1768 (2007).
[14] 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, No.542, pp.337–342 (2008).
[15] Crammer, K. and Singer, Y.: On the Learnability and Design of Output Codes for Multiclass Problems, *Machine Learning*, Vol.47, No.2-3, pp.201–233 (2002).
[16] 山口暢彦，石井直宏：エラー訂正符号を用いたパターン識別器の統合法，電子情報通信学会論文誌 D，Vol.86, No.2, pp.324–332 (2003).
[17] 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.104, No.706, pp.159–164 (2006).
[18] 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，pp.189–192 (2016).
[19] 平澤茂一，西島利尚：符号理論入門，培風館 (1999).
[20] 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, pp.2-1–2-2 (2017).
[21] 日外アソシエーツ：読売新聞データ集 (2015 年版) (2016).
[22] Lang, K.: Newsweeder: Learning to Filter Netnews, *Proc. 12th International Conference on Machine Learning*, pp.331–339 (1995).

雲居 玄道　（学生会員）

2008 年早稲田大学理工学部経営システム工学科卒業，2008 年同大学理工学術院総合研究所嘱託研究員．2015 年浄土真宗本願寺派総合研究所研究助手．2017 年早稲田大学大学院創造理工学研究科博士後期課程入学．2019 年早稲田大学創造理工学部経営システム工学科助手，現在に至る．情報数理応用・テキストマイニングの研究に従事．経営情報学会，日本気象学会各会員．

八木 秀樹

2001 年早稲田大学理工学部経営システム工学科卒業，2005 年同大学大学院理工学研究科博士後期課程修了．2005 年早稲田大学メディアネットワークセンター助手，2007 年同助教，2008 年電気通信大学先端領域研究センター特任助教，2008，2010 年プリンストン大客員研究員，2012 年電気通信大学大学院情報理工学研究科准教授，現在に至る．博士（工学）．情報理論，符号理論の研究に従事．IEEE，電子情報通信学会，信号処理学会各会員．

後藤 正幸　（正会員）

1994 年武蔵工業大学大学院修士課程修了．2000 年早稲田大学大学院理工学博士課程修了．博士（工学）．1997 年同大学理工学部助手．2000 年東京大学大学院工学研究科助手．2002 年武蔵工業大学環境情報学部助教授．2008 年早稲田大学創造理工学部経営システム工学科准教授．2011 年同大学教授．情報数理応用とデータサイエンスの研究に従事．著書に，『入門パターン認識と機械学習』コロナ社 (2014)，『ビジネス統計～統計基礎とエクセル分析』オデッセイコミュニケーションズ (2015) 等．IEEE，電子情報通信学会，人工知能学会，日本経営工学会，経営情報学会等各会員．

平澤 茂一　（正会員）

1961 年早稲田大学理工学部数学科卒業．1963 年同大学電気通信学科卒業．同年三菱電機（株）入社．1981 年早稲田大学理工学部工業経営学科（現，経営システム工学科）教授，2009 年早稲田大学名誉教授，早稲田大学理工学術院総合研究所研究員，現在に至る．情報理論とその応用，ならびに計算機応用システム等の研究に従事．工学博士（大阪大学）．1979 年 UCLA 計算機科学科客員研究員．1985 年ハンガリー科学アカデミー，1986 年イタリアトリエステ大学客員研究員，2002 年 UCLA 訪問教員．1993 年電子情報通信学会小林記念特別賞，業績賞受賞．IEEE Life Fellow，電子情報通信学会フェロー，経営情報学会等各会員．