F-011

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
A study of Construction method of error correction code in the multi-level document classification by ECOC method

雲居玄道* 小林学† 後藤正幸 * 平澤茂一 *
Gendo Kumoi Manabu Kobayashi 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 [2]．その代表的な手法として，Support Vector Machine(SVM) 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，もともと SVM は二値判別器であり，これを多値分類問題に適用する場合，1 つの分類器で直接モデル化する方法が考えられるが，計算量の問題で実用的とは言えない．この問題を回避するため，二値判別器を複数組み合わせて多値分類を実現可能であることが知られており，従来から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．その中の方法ひとつとして，符号理論の枠組みを導入した ECOC 復号法に基づく多値分類法がある [3].

ECOC 復号法に基づく多値分類法では，二値判別器の判定結果と各カテゴリにおける符号語とのハミング距離がそれぞれ計算され，この値が最も近いカテゴリへ復号するものである．この際，各カテゴリにどのような符号語を割り当てるかという点が ECOC 復号法に基づく多値分類法における重要な課題となっている．

本研究では，効率のよい符号語の構成について，実データによる評価実験から考察する．

2 準備

2.1 多値分類問題

分類問題とはカテゴリラベルの付いた入力を使って学習を行い，新たに与えられた入力 $\boldsymbol{x}$ に対応するカテゴリラベル $\mathcal{C} \in \{C_1, C_2, .., C_i, .., C_M\}$ を推定する問題のことである．M はカテゴリ数を表し，多値分類問題とは $M \geq 3$ の場合の分類問題のことを指す.

多値分類の手法としては，大きく分けて 2 通りのアプローチが存在する．1 つは多値分類問題を 1 つの分類器で直接モデル化するものである．もう 1 つの手法は複数の二値判別器の組み合わせで多値分類器を構成するものである．

2.2 Support Vector Machine

SVM[1] 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．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．

いま，入力ベクトルを x とする．このとき，各カテゴリ集合の学習データを分離する識別関数を，係数ベクトル $\boldsymbol{w}$，バイアス項 b を用いて，式 (1) で表す.

$$f(x) = \boldsymbol{w}^T \boldsymbol{x} + b \tag{1}$$

識別関数 $f(x)$ を求めるために，各学習データ x_i に対して，カテゴリ集合ラベルを $t_i \in \{-1, +1\}$，カテゴリ集合ラベル側のマージンからの誤差距離（スラック変数）を δi とし，最適化問題を解く.

2.3 ハミング距離

符号語の長さ（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，異なる個数をハミング距離を $D_H(\boldsymbol{a}, \boldsymbol{b})$ という．すなわち

$$D_H(\boldsymbol{a}, \boldsymbol{b}) = \sum_{n=1}^{N} d_H(a_n, b_n) \tag{2}$$

$$d_H(a_n, b_n) = \begin{cases} 0, & a_n = b_n \\ 1, & a_n \neq b_n. \end{cases} \tag{3}$$

である．さらに入力されたベクトル $\boldsymbol{y}$ に対し，符号語数を M，復号可能性のある符号語を $\boldsymbol{X} = \{x_1, x_2, \cdots, x_M\}$ とすると，

$$\boldsymbol{y} \to x_m; \arg\min_m D_H(\boldsymbol{y}, x_m) \tag{4}$$

とすると，$\boldsymbol{x}$ は x_m に復号する方法を最小距離復号法という.

2.3.1 符号語におけるハミング距離

符号理論において，符号語間の距離をハミング距離で計るが，その中でも最小のものを最小ハミング距離という．この最小ハミング距離を大きく取ることによって，訂正できる誤りが増える．具体的には，最小ハミング距離が d であったとき，$\lfloor d/2 \rfloor$ 個の誤りを訂正でき，$d-1$ 個の誤りを検出することができる．$\lfloor s \rfloor$ は，s を超えない最大の整数である．

2.3.2 符号語構成におけるハミング距離

符号語構成において，ある符号語間の距離が 1 ある場合，その符号語に割り当てられている 2 つのカテゴリを分類する分類器が 1 つ存在することを表している．つまり，ハミング距離が大きい符号語間には，そのカテゴリを分ける分類器多く含まれていることを表しており，2 つのカテゴリ同士において誤訂正を少なくすることができると言える．

3 ECOC 法に基づく多値分類法

誤り訂正 (ECC) 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 ECOC 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 [3].

*早稲田大学理工学術院
†湘南工科大学

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 W，W の (m,n) 要素 W_{mn} と表し，行列 W の各行の N 次元ベクトル $\boldsymbol{W}_m, (m = 1,2,\cdots,M)$ をカテゴリ C_m の符号語とする．

3.1 符号語構成法

符号語構成法は Dietterich と Bakiri が使っている Exhaustive Codes[3] をベースに用いる．分類器を $N_{MAX} = 2^{M-1} - 1$ 個作成する．

W_1 は全て 1 で構成する．W_2 は 2^{M-2} 個の 0 に続き $2^{M-2} - 1$ 個の 1 で構成する．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．W_i は 2^{M-i} 個の 0 と 1 を交互に並べて構成する．

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）が増大し，計算量が膨大となるため，分類に用いる分類器を減らし符号長を短くする必要がある．

3.2 判別方法

3.2.1 硬判定

判別方法は，符号語 $\boldsymbol{W}_{C_m}$ とテストデータに対する新規入力 $\boldsymbol{y}$ に対する M 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，

$$\hat{C} = \arg\min_{C_m} H_{C_m}, \tag{5}$$

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．

3.2.2 軟判定

SVM を用いた軟判定として，式 (1) 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 $\boldsymbol{W}_{C_m} = (a_1, a_2, \cdots, a_N)$ と入力 $\boldsymbol{x}$ に対する N 個の二値判別器の出力値 $f_{mn}(\boldsymbol{x})$ に対して，

$$G(\boldsymbol{W}_{C_m}, f_{mn}(\boldsymbol{y})) = \sum_{n=1}^{N} g(a_n) f_{mn}(\boldsymbol{y}) \tag{6}$$

$$g(a_n) = \begin{cases} 1, & a_n = 1 \\ -1, & a_n = 0. \end{cases} \tag{7}$$

とし，

$$\hat{C} = \arg\max_{C_m} G(\boldsymbol{W}_{C_m}, f_{mn}(\boldsymbol{y})), \tag{8}$$

とするカテゴリ $\hat{C}$ に判別する．

4 実験設定

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．

4.1 背景

本研究では，全てのカテゴリで学習データの数が等しいという問題設定のもとで，取りうることが可能な最長の符号長である Exhaustive Codes を基に分類器の数を減らすことにより符号長を縮め，効率のよい符号語構成法についての知見を得るため実験を行う．

4.2 符号語構成法

目的となる符号長の符号語を構成する場合には，Exhaustive Codes 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非復元抽出で選択し，構成を決定する．

4.3 判別方法

SVM を用いて硬判定を用いた場合，等距離となる符号語が存在する誤り検出可能であるが訂正不可能なケースが事前の実験より 3 割程度存在することが分かった．このため，本研究においては，このようなケースの存在しない SVM を用いた軟判定を用いて判別を行う．

4.4 実験方法

実験には，毎日新聞 2000 年の 8 カテゴリ (国際・経済・スポーツ・社会・芸能・家庭・総合・文化) の記事と読売新聞 2000 年 8 カテゴリ (政治・経済・スポーツ・社会・文化・生活・犯罪事件・科学) を使用する．すべての記事は 1 カテゴリだけに属し，カテゴリの重複はない．データから各カテゴリ 550 記事をランダムに選び，それを学習データ各カテゴリ 500 個，テストデータ 50 個にランダムに分ける．特徴量としては学習データに出現する全ての単語の単語頻度を使用する．カーネル関数は式 (9) で表される線形カーネルを用い，d は自然数であり，$d = 1$ とした．

$$K(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}^T \boldsymbol{y} + 1)^d. \tag{9}$$

5 実験結果

5.1 カテゴリ数 8 における実験

カテゴリ数 8 における Exhaustive Codes の符号長である 127 より，ランダムに分類器を選択する．実験においては，符号長 N を 7 から 127 まで 10 ずつ変化をさせて実験を行った．各符号長ごとにランダムに 1 万回，分類器を選択した．その結果を図 1 に示す．ただし，符号長 127 においては，符号長の最大値が 127 であることから，1 点のみとなる．

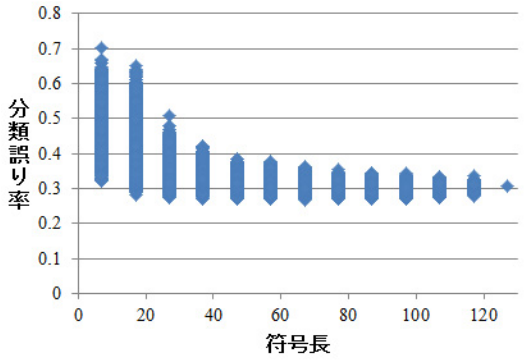

図 1. 各符号長における分類誤り率

表 1 に各符号長における分類誤り率の最小値および最大値を示した．図 1 より，Exhaustive Codes を全て用いることにより，分類誤り率を抑えられることがわかる．しかし，表 1 より，符号長が短くとも，Exhaustive Codes を全て用いた結果よりも分類誤り率 Pe を抑えることができる組合せが存在することが分かる．

表 1: 各符号長における分類誤り率

符号長	7	17	27	37	47	57	67	77	87	97	107	117	127
最小値	0.32	0.28	0.28	0.27	0.27	0.27	0.27	0.27	0.27	0.27	0.28	0.28	0.31
最大値	0.70	0.65	0.51	0.42	0.39	0.38	0.36	0.36	0.35	0.34	0.33	0.34	0.31

5.2 カテゴリ数 7 における実験

カテゴリ数 7 における Exhaustive Codes の符号長である 63 より，ランダムに分類器を選択する．実験においては，符号長を 7 から 127 まで 10 ずつ変化をさせて実験を行った．各符号長ごとにランダムに 1 万回，分類器を選択し，その平均値を図 2 に示す．ただし，カテゴリ数 8 より 7 へ減らす際に，減らす対象となるカテゴリは 8 通り存在するため，8 通りの実験を行った．図 2 のラベルは減らしたカテゴリの名称となる．

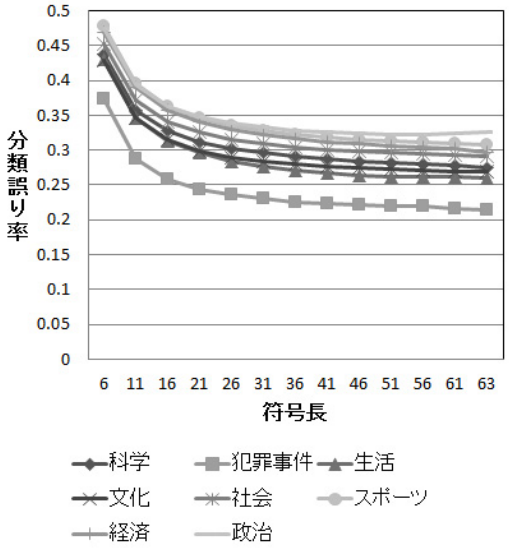

図 2. カテゴリ数 7 における分類誤り率

6 考察

図 1 より，符号長を長く取ることによって，分類器の性能のバラつきを抑えることができることが分かった．また，表 1 より，最大の符号長である 127 に対して，17 程度であっても，その性能を超える組合せが存在することを示した．このことより，分類器の選択方法を工夫することで，分類性能がよくかつ学習計算量を抑えた多値分類器を構成することが可能であることが分かる．

ここで，符号長 17 における最小の分類誤り率における符号語構成を表 2 に示す．また，符号語構成において，各符号語間のハミング距離を表 3 に示す．この結果から，「犯罪事件」のカテゴリと「政治」「経済」とのハミング距離が最大となっていることがわかる．図 2 の結果より，カテゴリ数を 7 と減らした場合において，「犯罪事件」のカテゴリを除いた場合に特に分類誤り率が低下することから，「犯罪事件」のカテゴリが最も分類することが困難なカテゴリということが推測される．このように，カテゴリ中に分類が困難なカテゴリが存在している場合，そのカテゴリに割当てる符号語を他のカテゴリの符号語に対してハミング距離をとることによって，分類性能を向上させられることが分かった．

これは，符号語の観点からみると，ハミング距離が大きいということは分類誤りに対して頑健であると言えると共に，分類器という観点からは，「犯罪事件」と「政治」や「経済」を分ける分類器が最も多く存在しており，誤判別を少なくすることができるていると言える．

ここで，符号長 27 についても同様の各符号語間の距離を表 4 に示す．この結果においても，最大のハミング距離は，「犯罪事件」と「経済」の間に存在している．

以上のことから，新聞記事といったデータにおいて，分類が困難なカテゴリが存在している場合に，そのカテゴリに対して他のカテゴリとの距離を大きくするような符号語構成が有効であることが示唆される．

7 まとめと今後の課題

本研究では，実際の文書分類問題に対し，SVM を用いた ECOC 法による多値文書分類における符号語構成について実験を行った．その結果，符号語構成の方法において，ハミング距離の有効性を知ることができた．

今後の課題は，有効な符号語構成について，対象となるデータに依らず，性能を保証できる構成方法を検討する必要がある．

参考文献

[1] V. Vapnik and A. Lerner, “Pattern recognition using generalized portrait method,” *Automation and Remote Control*, vol.24, pp. 774–780, 1963.

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

[3] 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

表 2: 符号長 17 における符号語構成

W_{C_1}	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
W_{C_2}	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
W_{C_3}	1	1	1	1	1	1	0	0	0	0	0	1	1	0	0	0	0
W_{C_4}	1	1	1	0	0	0	1	1	0	0	0	1	0	1	1	0	0
W_{C_5}	1	0	0	1	0	0	1	1	1	0	0	0	1	1	0	0	0
W_{C_6}	1	1	1	0	0	0	1	0	1	1	1	1	1	1	0	1	0
W_{C_7}	0	1	0	0	0	0	1	0	0	1	0	1	1	0	1	0	0
W_{C_8}	1	1	0	0	1	0	0	0	0	1	0	0	1	1	1	0	0

表 3: 各カテゴリカテゴリ間のハミング距離 ($N = 17$)

	政治	経済	スポーツ	社会	文化	生活	犯罪事件	科学
政治		6	9	9	10	6	**11**	10
経済	6		7	9	8	8	**11**	10
スポーツ	9	7		8	9	9	8	7
社会	9	9	8		7	7	6	7
文化	10	8	9	7		8	9	8
生活	6	8	9	7	8		7	8
犯罪事件	**11**	**11**	8	6	9	7		5
科学	10	10	7	7	8	8	5	

表 4: 各カテゴリカテゴリ間のハミング距離 ($N = 27$)

	政治	経済	スポーツ	社会	文化	生活	犯罪事件	科学
政治		15	12	12	12	12	16	17
経済	15		13	13	9	13	**19**	14
スポーツ	12	13		12	12	12	12	11
社会	12	13	12		18	16	12	11
文化	12	9	12	18		10	16	15
生活	12	13	12	16	10		16	13
犯罪事件	16	**19**	12	12	16	16		13
科学	17	14	11	11	15	13	13	