6B-01

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

雲居玄道 * 八木秀樹 † 後藤正幸 ‡ 平澤茂一 §
早稲田大学 * 電気通信大学 † 早稲田大学 ‡ 早稲田大学 §

1 はじめに

近年，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，文書分類などは複数のカテゴリを持つことも多く，一般には多値分類問題として定式化される．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，学習計算量はカテゴリ数 M に対して，$O(M^2)$ と膨大になる [3]．一方で，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．その中の方法のひとつとして，符号理論の枠組みを導入した ECOC（Error Correcting Output Codes）復号法に基づく多値分類法がある [4],[5]．この方法は二値判別器によって学習計算量を調整可能であり，最大の計算量は $O(2^M)$ であるが，最小では $O(M)$ となる．このことから，ECOC 法を用いてなるべく分類精度を高めつつ，計算量を抑えるような二値判別器の構成法が多値分類問題を解決する上で注目されている．

ECOC 復号法に基づく多値分類法では，誤り訂正符号の構造に着目し，二値判別器の判定結果を符号理論において伝搬されるビットの受信値とみなす．そして，二値判別器において生じる誤りを雑音のある通信路で生じた誤りと見なして，訂正し，判別結果の信頼度を上げている．

本稿では，二値判別器を用いた多値分類問題について，符号理論的な視点による効率のよい構成方法について考察し，ベンチマークデータを用い，その有効性を検証する．

2 誤り訂正符号 (ECC)[6]

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，情報を通信路に多少雑音が混入しても元の情報に訂正できるようにする手法である．

3 通信路と二値判別器

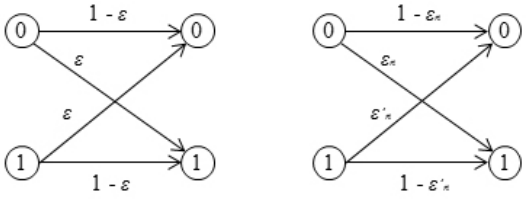
3.1 二元対称通信路

符号理論 [6] において，よく扱われる通信路のモデルとして，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．

3.2 二値判別器

Dietterich と Bakiri [4] 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 [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 $\tilde{\boldsymbol{w}}_n$ $(n = 1,2,\cdots,N)$ を二値判別器 n のカテゴリ分割ベクトルとする．

この時，二値判別器（各ビット）においては，N 個の二値判別器（ビット位置 n）ごとに誤判別率（雑音発生確率）が異なる（図 2）．すなわち，$0 \to 1, 1 \to 0$ と誤る確率が異なることから，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

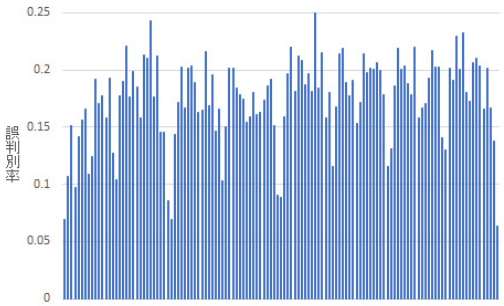


図 2. ECOC 法における二値判別器ごとの誤判別率 ($M = 8, N = 127$ の例)

4 符号語構成法

4.1 Exhaustive Codes

Exhaustive Codes [4] は M 個のカテゴリを 2 つに分割する冗長性を除いた全ての組み合わせに対応する二値判別器を用意する構成法であり，二値判別器を $N_{MAX} = 2^{M-1} - 1$ 個作成する．すなわち分類器の個数（最大符号長）はカテゴリ数 M に応じて指数的に増大し，計算量が膨大となるため，効率的な構成法を検討する必要がある．

A study on construction of ECOC matrix for multi-valued document classification based on performance of binary classifiers

*Gendo Kumoi,Waseda University
†Yagi Hideki,The University of Electro-Communications
‡Masayuki Goto,Waseda University
§Shigeichi Hirasawa,Waseda University

¶ただし，情報系列とパリティ系列が区別できない非組織符号も含めて考える

4.2 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，カテゴリが既知の D 個の文書は，それぞれカテゴリ C_{c_d} に対して，$\boldsymbol{w}_{c_d}$ 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 $\hat{\boldsymbol{W}}_{c_d} = \{\hat{w}_{c_d1}, \hat{w}_{c_d2}, \cdots, \hat{w}_{c_dN}\}$ と受信すると見なせる．

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 $\tilde{w}_n, \hat{w}_n$ と書く．この時，二値判別器の平均誤判別率は，

$$\begin{aligned} \bar{P}_n &= \Pr\{\tilde{w}_n = 0\}\Pr\{\hat{w}_n = 1|\tilde{w}_n = 0\} \\ &+ \Pr\{\tilde{w}_n = 1\}\Pr\{\hat{w}_n = 0|\tilde{w}_n = 1\} \end{aligned} \quad (1)$$

と表すことができる．これは，符号理論におけるビットエラーレート（BER）と捉えることができ，これらは小さい方が信頼度が高い通信路と見なせる．

4.3 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．判別器 n の相互情報量は，

$$\begin{aligned} I_n &= \Sigma_{y=0}^{1}\Sigma_{\hat{y}=0}^{1}\Pr\{\tilde{w}_n = y\}\Pr\{\hat{w}_n = \hat{y}|\tilde{w}_n = y\} \\ &\times \log \frac{\Pr\{\hat{w}_n = \hat{y}|\tilde{w}_n = y\}}{\Sigma_{y'=0}^{1}\Pr\{\tilde{w}_n = y'\}\Pr\{\hat{w}_n = \hat{y}|\tilde{w}_n = y'\}} \end{aligned} \quad (2)$$

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 性能評価

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 を基にランダムに二値判別器を選択する（ランダム）構成法と誤判別率，相互情報量に基いて選択した構成法を比較する．

5.1 判別方法

符号理論においては受信系列を復号する手法のひとつに最小距離復号がある．ECOC 法においては，送信した符号語との距離を SVM を用いた軟判定値 [7] を用いて算出し，カテゴリを推定する [4].

5.2 実験方法

5.2.1 実験データ

実験には，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) の記事 [8] 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 (3,059 件) を使用する．カーネル関数は線形カーネルを用いた．

5.2.2 符号語構成

Exhaustive Codes に対して，8 カテゴリにおける最小符号長を 7 とし，10 刻みに 127 まで 10,000 回ずつ実験を行った．この際，

1. Exhaustive Codes から，ランダムに符号長に対して二値判別器を選択した（ランダム構成法）.
2. 誤判別率に対しては，Exhaustive Codes より，$\bar{P}_n$ が小さいものから順に選択した（誤判別率構成法）.
3. 相互情報量に対しては，Exhaustive Codes より，I_n が大きいものから順に選択した（相互情報量構成法）.

ここで，誤判別率，相互情報量については，学習データから二値判別器を学習した後に，同じ学習データに対して，分類を行うことにより，それぞれの値を推定した．

6 実験結果

ランダム，誤判別率，相互情報量による構成の結果を図 2 に示す．

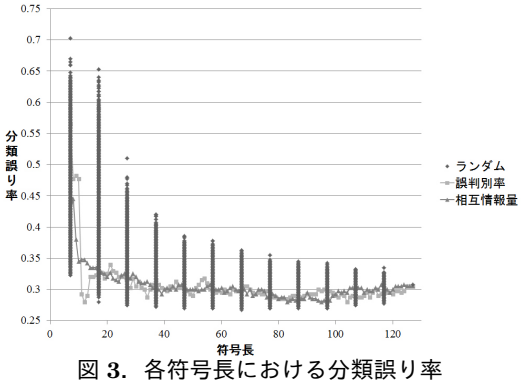


図 3. 各符号長における分類誤り率

図 3. よりランダムに選択した場合，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 点のみとなり，符号長が短い場合にはばらつきが大きくなった．誤判別率，相互情報量による構成は，どちらもそのばらつきに対して，分類誤り率が平均よりも下回る結果を得ることができた．このことから，誤判別率，相互情報量に着目し符号語表を構成することは有効であると考えられる．

また，符号長 12 の誤判別率における符号語構成において，最小の分類誤り率を達成した．これは，12 個の二値判別器が 1 vs The Rest と呼ばれる．対象となる 1 カテゴリ対その他の 7 カテゴリとして二値判別器を学習したものが 8 個，揃ったところである．このことから，規則的な構成を符号語構成に含めることも分類誤り率に対して有効であると考えられる．

7 まとめと今後の課題

本研究では，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，符号表の構成において，誤判別率，相互情報量に着目することの有効性を示した．

今後は，対象となるデータに依らず，性能を保証できる符号表の構成方法を検討する必要がある．

謝辞 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，挑戦的萌芽研究 26560167 の助成による．

参考文献

[1] B. 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.

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

[3] J. Weston and C. Watkins, “Support vector machinesdor multi-class pattern recognition,” *Proceedings ESANN'99*, pp.219–224, 1999.

[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] 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.

[6] 平澤茂一，西島利尚，“符号理論入門，” 培風館,1999.

[7] 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.

[8] 読売新聞記事データ 2000 年版, 日外アソシエーツ株式会社.