

2値 ECOC と 3値 ECOC のシステム評価

System Evaluation of Binary ECOCs and Ternary ECOCs

平澤茂一¹ 雲居玄道¹ 八木秀樹² 小林学¹ 後藤正幸¹ 稲積宏誠³
 Shigeichi Hirasawa Gendo Kumoi Hideki Yagi Manabu Kobayashi Masayuki Goto Hiroshige Inazumi

早稲田大学¹ 電気通信大学² 青山学院大学³
 Waseda University University of Electro-Communication Aoyama Gakuinn University

1 まえがき

ECOC (Error-Correcting Output Code) は $N(\geq \lceil \log_2 M \rceil)$ 個の 2 値 ($\{0, 1\}$) 判別器を用いて, $M(\geq 3)$ 個のカテゴリを持つ多値分類システムを構成する有力な方法の 1 つである [1]. 2 値 ECOC は 3 値 ($\{0, *, 1\}$) ECOC にも拡張されている. ただし, $*$ は don't care である [2]. いずれも設計が容易で分類性能に優れ, 多くの研究がある.

一方, 情報システムの性能劣化 (d) と投資コスト (ℓ) のトレードオフ関係をモデル化したシステム評価モデルが, 1970 年代後半にレート歪理論に基づいて J. Pearl らによって提案され [3], 後に筆者らにより適用を容易にしたシステム評価モデルに拡張されている [4]. なお, このトレードオフモデルは, 「ほんの僅かな性能劣化を許容すれば, コストを大幅に減少出来る」こと (Flexible 性) [3] に注目している.

ここでは, 2 値および 3 値 ECOC に上記のシステム評価モデルを適用し, 優れた性質を持つことを示す.

2 Exhaustive 符号と短縮 Exhaustive 符号

カテゴリ数 M が与えられたとき, 分類に寄与するすべての列ベクトルを選んで生成される符号を Exhaustive 符号と呼び, 符号長 $N_{\max}$ は

$$2 \text{ 値} : N_{\max} = 2^{M-1} - 1, \quad (1)$$

$$3 \text{ 値} : N_{\max} = \sum_{\alpha=0}^{M-2} \binom{M}{M-\alpha} (2^{M-1-\alpha} - 1), \quad (2)$$

与えられる [1][2]. 表 1 に 3 値 Exhaustive 符号 ($M=4, N_{\max}=25$) の例を示す. また, $N_{\max}$ から $N(N_{\min} \leq N \leq N_{\max})$ 個の列ベクトルを選んで構成される符号を短縮 Exhaustive 符号と呼ぶことにする. ただし, $N_{\min} = \lceil \log_2 M \rceil$ とする.

表 1 3 値 Exhaustive 符号 ($M=4, N_{\max}=25$)

	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}	d_{16}	d_{17}	d_{18}	d_{19}	d_{20}	d_{21}	d_{22}	d_{23}	d_{24}	d_{25}
c_1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	*	*	*	1	1	1	*	*	*
c_2	0	0	0	0	1	1	1	0	1	1	0	1	1	*	*	*	0	0	0	0	*	*	1	1	*
c_3	0	0	1	1	0	0	1	1	0	1	*	*	*	0	1	1	0	1	1	*	0	*	0	*	1
c_4	0	1	0	1	0	1	0	*	*	*	1	0	1	1	0	1	1	0	1	*	*	0	*	0	0

3 実験

(1) 実験データ: 2015 年読売新聞記事 [5] から 8 個のカテゴリを用い, 各カテゴリあたり 1,500 件の文書 (全データ数は 12,000) に対して 10 分割交差検証により分類誤り率を求める.

(2) 実験方法と結果: 8 個のカテゴリからランダムに M 個のカテゴリを選び, $N_{\max}$ 個の列ベクトルからランダムに N 個を選んで構成される短縮 Exhaustive 符号の平均分類誤り率 (P_e) と符号長 (N) のトレードオフ曲線 $P_e = S(N, M)$ を実験的に求める. さらに, $p_e = P_e/P_{e,\max}$, $n = N/N_{\max}$ とし

$$p_e = s(n, M) \quad (3)$$

を求める. ここで, $P_{e,\max} = S(N_{\min}, M)$ である. 結果を図 1 に示す.

4 考察とむすび

$M=4, 5, 6, 7$ に対し, 図 1 よりトレードオフ曲線から

1. 全ての ECOC は Flexible である.
2. 与えられた M に対し, 3 値 ECOC の方が 2 値 ECOC より, More Flexible である.
3. $n=1$ の近辺を除き, 2 値 ECOC と 3 値 ECOC ともに Elastic である.
4. 2 値 ECOC と 3 値 ECOC ともに Effective Elastic である (詳細は [6] 参照).

以上より ECOC は情報システムの要素として, 特に M が大きくなると望ましい性質を持っており, さらに深く詳細に検討する価値がある.

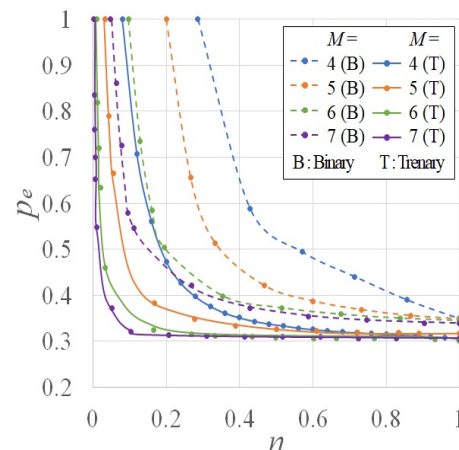


図 1 投資コスト割合 n と性能の劣化割合 p_e のトレードオフ関係 (パラメータ: システムの規模 M).

謝辞 本研究の一部は学振研究助成基金助成金 (C)19K04914, (B)19H01721, (C)18K11585 の助成による.

参考文献

- [1] T. G. Dietterich and G. Bakiri, "Solving multi-class learning problems via error-correcting output codes," *Journal of Artificial Intelligence Research*, vol. 2, pp. 263–286, 1995.
- [2] S.-H. Park, and J. Frnkranz, "Efficient decoding of ternary error-correcting output codes for multiclass classification," *Technical Report TUD-KE-2009-01*, Technische Universitt Darmstadt, pp. 1–20, 2009.
- [3] 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.
- [4] S. Hirasawa, and H. Inazumi, "A system evaluation model by using information theory," *The 30th Joint National Meeting, ORSA/TIMS*, MB35.3, Philadelphia, PE, USA, Oct. 1990.
- [5] Yomiuri News Paper Articles 2015, *Nichigai Associates Inc.*
- [6] S. Hirasawa, G. Kumoi, H. Yagi, M. Kobayashi, M. Goto, T. Sakai, and H. Inazumi, "System evaluation of ternary error-correcting output codes for multiclass classification problems," (accepted for presentation at the 2019 IEEE Int. Conf. on SMCC).