3 元直交表に基づく ECOC 法による多値文書分類

1X11C066-5　鈴木 玲央奈
指導教員　後藤 正幸

1　研究背景・目的

近年，重要性が増している自動文書分類の多くは分類先のカテゴリ数が 3 以上の場合を対象とした多値判別問題である．多値判別問題の解法には，様々なアプローチが存在するが，本研究では強力な二値判別器を組み合わせる，Error-Correcting Output Coding 多値判別法 [1](以下，ECOC 法) に注目する．ECOC 法は，各行にカテゴリ，各列に二値判別器を表した符号表と呼ばれる 0,1 の数値表を用いて判別器の構成を表現し，二値判別器の出力結果から新規文書の所属カテゴリを推定するものである．一方，ECOC 法では各二値判別器で分類する 2 つのカテゴリ集合間で学習データのバランスが悪くなると予測精度が劣化することが知られている．この学習データ数の比を考慮した手法として Reed Muller 符号（以下 RM 符号）を用いる手法がある [2]．しかしながら，RM 符号を用いた手法は，特定のカテゴリ数に対してしかカテゴリ集合間で学習データ数を等しくすることができず，任意のカテゴリ数に適用可能な方法が望まれる．

本研究では，判別に用いないカテゴリが存在する二値判別器を用いることにより，任意のカテゴリ数に対して，2 つのカテゴリ集合間で学習データ数の比を 1 : 1 に近づける手法を提案する．結果，学習に用いるデータ数が減り，計算量の低減も可能となる．提案手法を新聞記事データを用いた文書分類問題に適用し，精度・計算量の面での有効性を示す．

2　準備

2.1　ECOC 法

ECOC 法とは，符号理論で用いられる誤り訂正技術を自動分類に応用した手法であり，カテゴリが未知の入力データに対し，複数の二値判別器を組み合わせて所属カテゴリ $c_k(k = 1, ..., K)$ を推定する．K はカテゴリ数である．そのための各二値判別器の構成を符号表と呼ばれる {0,1} の二値で表される数値表により表現する．いま，符号表を $\boldsymbol{W}$，二値判別器の個数を R とすると，$\boldsymbol{W}$ は $K \times R$ 行列となる．符号表 $\boldsymbol{W}$ の各列ベクトルは二値判別器の構成を表現しており，要素が 1 のカテゴリ集合と要素が 0 のカテゴリ集合を二値判別するという意味になる．そのため，ある 2 つの判別器を表す列ベクトル同士が互いに 0 と 1 が反転しただけの場合，その 2 つは等価な判別器となる．また，符号表 $\boldsymbol{W}$ の i 行目の行ベクトルをカテゴリ c_i の符号語と呼び $\boldsymbol{W}_{c_i}$ と表現する．

本研究では，二値判別器として判別精度が高く，出力値がカテゴリへの所属確率となる Relevance Vector Machine（RVM）[3] を用いる．$r(r = 1, ..., R)$ 番目の判別器の出力を G_r，符号語 $\boldsymbol{W}_{c_i}$ の r 番目の値を $W_{c_i \cdot r}$ とすると，新規入力に対する分類は以下の式で行う．

$$\hat{c} = \arg\max_{c_i} \prod_{r=1}^{R} G_r^{W_{c_i \cdot r}} (1 - G_r)^{W_{c_i \cdot r}} \qquad (1)$$

符号表の中には {1,0} の二値で表される 2 元符号表のほかに，判別に用いないカテゴリを許容した 3 元符号表がある．ここでは，判別に用いないカテゴリを $*$ で表し，$W_{c_i \cdot r}$ が $*$ の場合には判別器 r においてカテゴリ c_i の学習データは用いない．そのため，3 元符号表は各二値判別器で用いる学習データ数が減り，2 元符号表よりも学習の計算量を低減させることができる．

他方，ECOC 法において分類精度が高くなる条件として，符号語 $\boldsymbol{W}_{c_i}$ 間の最小ハミング距離が大きいことがあげられる．これは符号語間の距離を離すことでカテゴリを識別しやすくするためである．一般に判別器数が多い符号表を用いた場合，符号語間のハミング距離を大きくすることができるため，分類精度も向上するが，その分，計算量も増加する．

2.2　Reed Muller 符号を用いた ECOC 法

RM 符号とは，符号理論で用いられる符号の 1 つであり，ECOC 法との相性が良いことが知られている．RM 符号を用いるとカテゴリ数が 2 の冪乗の場合は，符号表の各列の 0 と 1 の数が $K/2$ ずつになる．そのため，各カテゴリの学習データ数が等しい場合，カテゴリ集合間で学習データ数が等しくなる．加えて,「各カテゴリの符号語間のハミング距離が大きい」,「各判別器間のハミング距離が全て等しい」という特徴もある．これらの 3 つの特徴から，カテゴリ数が 2 の冪乗の場合には，RM 符号による ECOC 法により分類精度の高い判別器構成が可能である．

3　提案手法

3.1　提案手法の概要

RM 符号による ECOC 法では，K 個のカテゴリを 2 つのカテゴリ集合に等分する．そのため，K が奇数の場合，カテゴリ集合間で学習データ数が等しくなるような判別器を構成できない．この問題は全カテゴリを 3 つのカテゴリ集合に分割し，そのうちの 2 つのカテゴリ集合間で学習データ数が等しくなるようにすることで解決可能である．学習データ数が等しい 2 つのカテゴリ集合間で二値判別を行い，残りのカテゴリ集合は学習に用いない 3 元符号表を作成する．

3.2　探索的 3 元符号表の作成

全カテゴリを 3 つのカテゴリ集合に 3 等分し，そのうち 2 つのカテゴリ集合間で二値判別を行うことを考える．全カテゴリを 3 等分する符号表は，各列において {0,1,2} の 3 つの元の数が等しい表を用いることで作成できる．例えば実験計画法で用いられる 3 元直交表はこれを満たす．

提案手法では，3 元直交表を初期行列とし，{0,1,2} の 3 つの元の数が等しい列ベクトルを順次新たな列として追加したものを符号表とすることで，判別器数を増加させ，分類精度の向上を行う．このとき単純に {0,1,2} の 3 つの元の数が等しい任意のベクトルを追加した場合,「符号語間の距離が小さくなってしまう」,「類似した判別器ができてしまう」という 2 つの問題点が発生する場合がある．提案手法ではこれを避けるために，前者の問題点に対しては，最も符号語が類似しているカテゴリ，すなわち行間のハミング距離が最小となるカテゴリの組に着目し，その最小ハミング距離を最も大きくするベクトルを選択する．後者の問題点に対しては，追加候補のベクトルと各列とのハミング距離の最大値と最小値に着目する．追加候補のベクトルと類似した列が存在する場合，最大値か最小値，またはその両方が極端な値になることから，提案手法では最大値と最小値の差が最も小さいベクトルを各列とのハミング距離が極端でないベクトルとして選択する．これは，ECOC 法では 0 と 1 が反転しただけの判別器構成は等価となり，類似した判別器は判別器間のハミング距離が極端に大きい，または小さくなるためである．

さらに直交表に列を追加することで得られた各列において {0,1,2} の 3 つの元の数が等しい表を ECOC 法における {0,1,*} の 3 元符号表へと変換を行う．変換の方法は {0,1,2} のどれか 1 つの元を $*$ にして，残り 2 つの元に 0 と 1 を割り当てる．そのため，{0,1,2} のどの元を $*$ にするかの 3 通り

の変換が可能であり，提案手法ではその 3 通り全てを行い，得られる 3 つの符号表を結合し 1 つの符号表として用いる．

3.3 符号表作成アルゴリズム

提案アルゴリズムの具体的な流れを以下に述べる．まず，K 行以上の $\{0,1,2\}$ の 3 元直交表から行数が最小である 3 元直交表を作成する．K_H を作成した直交表の行数とする．次にその直交表に「行間のハミング距離の最小値を大きくする」,「各列とのハミング距離の最大値と最小値の差が最小となる」という 2 つの基準を用いて列ベクトルを追加する．追加するベクトルは K_H 次元で各成分が {0,1,2} からなり，各元の数が $K_H/3$ ずつとなるベクトルである．閾値をこえるまで列追加を繰り返した後，得られた $\{0,1,2\}$ の 3 元の表を {0,1,∗} の ECOC 法に適用可能な 3 元符号表へと変換を行う．以下では，α を列追加のための閾値，$d_H(\boldsymbol{x},\boldsymbol{y})$ を $\boldsymbol{x}$ と $\boldsymbol{y}$ のハミング距離とする．

Step1) K_H 行 N_H 列の 3 元直交表を $\boldsymbol{H}$ とし，$\boldsymbol{H}$ の各列ベクトルを $\boldsymbol{h}_i(i = 1,...,N_H)$，各行ベクトルを $\boldsymbol{w}_k(k = 1,...,K_H)$ とする．

Step2) K_H 次元で各元 $\{0,1,2\}$ の数が $K_H/3$ ずつとなる考えられる全てのベクトル N 個を追加する列の候補とし，$j(j = 1,...,N)$ 番目のベクトルを $\boldsymbol{h}'_j = (h'_{j,1},...,h'_{j,K_H})^T$ とする．

Step3) 各 j について，全ての i に対して $d_H(\boldsymbol{h}_i,\boldsymbol{h}'_j)$ を計算し，$d_H(\boldsymbol{h}_i,\boldsymbol{h}'_j)$ の最大値と最小値の差を $R(\boldsymbol{h}'_j)$ とする．

Step4) $R(\boldsymbol{h}'_j)$ の最小値を R_{min} とする．$R_{min} \leq \alpha$ なら Step5 へ．さもなければ Step8 へ．

Step5) $\boldsymbol{H}$ の k_1,k_2 行間のハミング距離，$d_H(\boldsymbol{w}_{k_1},\boldsymbol{w}_{k_2})$ を全ての k_1,k_2 の組み合わせに対して計算する．

Step6) 全ての k_1,k_2 の組み合わせに対して，$d_H(\boldsymbol{w}_{k_1},\boldsymbol{w}_{k_2})$ が最小となる k_1,k_2 を k_1^{min},k_2^{min} とする．複数存在する場合はランダムに選択する．

Step7) $N_H = N_H + 1$ とする．全ての $\boldsymbol{h}'_j$ に対して，$R(\boldsymbol{h}'_j) = R_{min}$，かつ $h'_{j,k_1^{min}} \neq h'_{j,k_2^{min}}$ となる $\boldsymbol{h}'_j$ を $\boldsymbol{h}_{N_H}$ として $\boldsymbol{H}$ の列に追加する．複数存在する場合はランダム選択して追加し Step3 へ戻る．

Step8) 行列 $\boldsymbol{H}$ に関して，$\{0,1,2\}$ それぞれの元に対して $\{0\to0, 1\to1, 2\to*\}$ としたものを $\boldsymbol{H}_1$，$\{0\to1, 1\to*, 2\to0\}$ としたものを $\boldsymbol{H}_2$,$\{0\to*, 1\to0, 2\to1\}$ としたものを $\boldsymbol{H}_3$ とし，$[\boldsymbol{H}_1,\boldsymbol{H}_2,\boldsymbol{H}_3]$ と結合し，符号表とする． □

上記，Step3 では全ての列ベクトルに対し，$\boldsymbol{H}$ の列ベクトルとのハミング距離の範囲を表す $R(\boldsymbol{h}'_j)$ を計算し，Step4 で $R(\boldsymbol{h}'_j)$ を最小とするベクトルを列に追加する．Step5,6 で $\boldsymbol{H}$ の各行間のハミング距離が最小の行の組を選択し，Step7 で Step4 で抽出したベクトルからハミング距離最小の行間のハミング距離を大きくするベクトルを選択し，列に追加する．

表 1. 提案アルゴリズムにより得られる符号表の例

	01	02	03	⋯	18	19	20	⋯	50	51
c_1	0	0	0	⋯	1	1	1	⋯	∗	∗
c_2	0	1	1	⋯	1	∗	∗	⋯	0	0
c_3	0	∗	∗	⋯	1	0	0	⋯	1	1
c_4	1	0	1	⋯	∗	∗	∗	⋯	∗	0
c_5	1	1	∗	⋯	∗	0	0	⋯	0	∗
c_6	1	∗	0	⋯	∗	1	1	⋯	0	0
c_7	∗	0	∗	⋯	0	0	0	⋯	1	∗
c_8	∗	1	0	⋯	0	1	1	⋯	∗	1
c_9	∗	∗	1	⋯	0	∗	∗	⋯	1	1

表 1 は $K = 9$ の場合の提案手法によって得られた符号表の例である．各列において 0 と 1 の数が等しく，直交表に比べて判別器数が増加している．また各判別器における学習データ数が少なく，全体の計算量を抑えることもできている．また，$K < K_H$ の場合には，上記のアルゴリズムにより作成された符号表から K 行選択し，適用する．

4 実験・考察

提案手法の有効性を検証するために新聞記事を用いた分類実験を行った．実験データは 2000 年の毎日新聞記事から 9 カテゴリを使用した．学習データは各カテゴリ 100 件，200 件とし，テストデータは各カテゴリ 100 件とした．評価指標はテストデータに対する正解率，学習の計算時間の 2 つとし，それぞれ 6 回の実験の平均を用いた．比較手法は，Exhaustive 符号 [1]，RM 符号，3 元直交表を 3 元符号表へと変換したものを用い，提案手法は 3 元直交表に列を追加したものとする．比較手法である Exhaustive 符号は考えられる全ての 2 元の判別器を用いる手法であるため高い分類精度となるが，判別器数が多くなり計算時間が増加することが知られている．また，RM 符号は直接 9 カテゴリに適用できないため，16 カテゴリの RM 符号からランダムに 9 行選択することを 10 回行い，それらの平均を用いた．提案手法は直交表に列を追加するときに候補が複数ある場合はランダムに追加するため，10 通りの結果の平均を用いた．提案手法における列追加のための閾値は予備実験により，$\alpha = 3$ とした．各比較手法の判別器数を表 2 に，実験結果を表 3 に示す．

表 2. 各手法の判別器数

符号表	判別器数
Exhaustive 符号	255（2 元）
RM 符号	15（2 元）
直交表	12（3 元）
提案手法（平均）	47.4（3 元）

表 3. 実験結果

符号表	学習データ数 100		学習データ数 200	
	正解率	計算時間 (秒)	正解率	計算時間 (秒)
Exhaustive 符号	0.682	13,514	0.731	94,094
RM 符号	0.669	835	0.723	5,754
直交表	0.651	214	0.710	1,329
提案手法	0.678	1,125	0.733	7,771

表 3 より，正解率は学習データ数 100 の場合は Exhaustive 符号，学習データ数 200 の場合では提案手法が最大となった．計算時間はどちらの学習データ数の場合でも，判別器数順となった．この結果より，提案手法は精度を十分に保ったまま計算量を削減できているといえる．

直交表のみを用いた手法は，判別器数は RM 符号と大きな差はないが，RM 符号より低い正解率となった．これは各二値判別器における学習データ数が 3 元の判別器を用いたために低下していることが原因だと思われる．一方で，学習データ数が少ないことにより極端に少ない計算時間となっていることから，提案手法において列追加の閾値 $\alpha = 3$ を小さくすることで，RM 符号と同等の正解率で，少ない計算時間となることも期待できる．

5 まとめと今後の課題

本研究では，文書分類問題を対象とし，RM 符号ではカテゴリ集合間で学習データ数を等しくできなかったカテゴリ数に対して，そのアンバランスを解消できる符号表の作成アルゴリズムを提案した．実験結果より，計算量の増加を抑えつつ高い正解率となることを示した．今後の課題として，3 元以外の直交表の適用，カテゴリ数の拡張の提案が挙げられる．

参考文献

[1] T. G. Dietterich and G.Bakiri. “Solving Multiclass L.earning Problems via Error-Correcting Output Codes,” *Artif. Intell.,* vol.2, pp. 263-286, 1995.

[2] 荻原大陸，三川健太，後藤正幸，“Reed Muller 符号を用いた階層的 ECOC 法による多値文書分類,” 第 36 回情報理論とその応用シンポジウム，pp. 352-357, 2013.

[3] M. E. Tipping. “Sparse Bayesian Learning and the Relevance Vector Machine,” *Mach. Learn. Res.,* pp. 211-244, 2001.