E-017

サブカテゴリを用いた ECOC 法による多値文書分類に関する一考察
Multi-Category Document Classification Based on ECOC Approach Using Sub-categories

鈴木 玲央奈 † 山下 遥 † 後藤 正幸 †
Leona Suzuki Haruka Yamashita Masayuki Goto

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 3 カテゴリ以上の多値判別問題である．本研究では，様々な多値判別手法の中でも，強力な二値判別器を組み合わせる ECOC 法 [1] に着目する． ECOC 法は，各行にカテゴリ，各列に二値判別器を対応させた符号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データの所属カテゴリを推定するものである．本研究では，分類精度の向上のために，潜在的なサブカテゴリを積極的に用いた二値分類器集合の新たな構成法とこれらの二値分類器集合を用いた新たな多値分類法を提案する．提案手法の有効性について新聞記事データを用いた文書分類問題に適用し，検証を行う．

2 ECOC 法

2.1 ECOC 法概要

ECOC 法は，符号理論で用いられる誤り訂正技術を多値判別問題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，所属カテゴリ $c_k(1 \leq k \leq K)$ を推定する．複数の二値判別器の構成は符号表と呼ばれる {1,0} の二値で表される数値表により表現する．いま，二値判別器の個数を R とすると，符号表 $\boldsymbol{W}$ は $K \times R$ 行列となる．すなわち，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．そのため，0 と 1 が反転した列ベクトルは，等価な判別器を表す．また，符号表 $\boldsymbol{W}$ の k 行目の行ベクトルをカテゴリ c_k の符号語と呼び $\boldsymbol{W}_k$ と表現する．新規データの所属カテゴリを推定する際には，新規データに対する各二値判別器の出力結果ベクトルと各カテゴリの符号語を比較して，分類を行う．

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，$\boldsymbol{W}_k$ の r 番目の値を W_k^r とすると，W_k^r が $*$ の場合には判別器 r においてカテゴリ k の学習データは用いないものとする．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 元符号表の場合よりも学習計算量が低減する．

2.2 従来の符号表

代表的な符号構成法として，一対他法と Exhaustive 符号 [1] がある．一対他法とは，1 つのカテゴリとそれ以外を二値判別する判別器をカテゴリ数分用いる符号表構成であり，$R = K$ となる．Exhasutive 符号とは，Dietterich らによって示された符号表であり，考えられる全ての 2 群分類に対する判別器を用意する判別器構成となっている．そのため高い分類精度となる一方で，判別器数が $2^{K-1} - 1$ と膨大になるため計算時間も膨大となってしまう．

2.3 類似度最大に基づく分類基準

入力 $\boldsymbol{x}$ に対する $r(1 \leq r \leq R)$ 番目の二値判別器の出力を $G_r(\boldsymbol{x})$ としたとき，類似度最大に基づく分類基準では，符号語 $\boldsymbol{W}_k$ と $\boldsymbol{G} = (G_1(\boldsymbol{x}), ..., G_R(\boldsymbol{x}))$ の類似度 $S(\boldsymbol{W}_k, \boldsymbol{G})$ が最大となるカテゴリを式 (1) により導出し，分類する．

$$\hat{k} = \arg\max_k S(\boldsymbol{W}_k, \boldsymbol{G}) \quad (1)$$

すなわち，ECOC 法の分類ルールは，二値分類器の出力例から，最も類似性の高い唯一の符号語を選ぶ操作となっている．

3 提案手法

3.1 提案手法の着眼点

提案手法では，符号表生成と分類基準の双方に着目する．まず，符号表生成に関しては，サブカテゴリを用いた符号表生成を考える．従来の符号表ではカテゴリ情報のみを用いて符号表の生成が行われている．しかしながら多くのデータにおいて，同一カテゴリ内において性質が異なる複数のサブカテゴリが存在する場合や，異なるカテゴリに属するサブカテゴリ同士の特徴が類似する場合が考えられる．そこで，カテゴリ情報に加え，サブカテゴリの特徴を考慮した符号表の生成が望まれる．そこで提案手法では，カテゴリをデータの特徴からいくつかのサブカテゴリへと分割し，そのサブカテゴリを用いた符号表を生成することにより，これを達成する．ここでカテゴリ c_k から分割された j 番目のサブカテゴリを $c_{k,j}$ とすると，サブカテゴリを用いた符号表は表 1 のように表せる．このように，サブカテゴリに対し 1 つの符号語を対応させることで，同一カテゴリ内においても，異なるカテゴリ集合とみなして二値判別を行う判別器構成を構成することが可能となる．

表 1: サブカテゴリを用いた符号表例

		判別器				
c_k	$c_{k,j}$	1	2	3	⋯	R
c_1	$c_{1,1}$	1	0	0		1
	$c_{1,2}$	0	0	1		1
	$c_{1,3}$	0	1	0		1
c_2	$c_{2,1}$	1	0	0	⋯	0
	$c_{2,2}$	0	1	0		0
c_3	$c_{3,1}$	0	0	1		0
	$c_{3,2}$	0	1	0		0

また，分類基準に関しては，各カテゴリが複数の符号語を持つ場合に適した分類ルールを考える．従来の類似度最大に基づく分類基準では，各カテゴリにおいて 1 つの符号語しか存在しないことを前提とした分類基準である．しかしながら，サブカテゴリを用いた符号表では，表 1 のように各カテゴリが複数の符号語を持つ．そこで，提案手法では，各カテゴリが複数の符号語を持つ場合にノイズに対して頑健となる分類ルールを考える．

3.2 符号表生成アルゴリズム

この節では，提案する符号表生成アルゴリズムについて述べる．符号表生成アルゴリズムの大まかな流れは，1) サブカテゴリのランダム生成，2) 二値判別器構成の決定（$0, 1, *$ の割り当て）になる．まず，サブカテゴリの生成では，各カテゴリごとにサブカテゴリの代表ベクトルをランダムに選択し，代表ベクトルと学習データ間の距離による学習データのグルーピングにより，サブカテゴリを生成する．これにより，特徴の異なるいくつかのサブカテゴリにカテゴリを分割可能

†早稲田大学

となる．また，2 つ目の手順の二値判別器構成の生成では，得られたサブカテゴリをサブカテゴリ間の距離によりグルーピングを行い，サブカテゴリグループを生成する．生成されたグループをカテゴリとみなし $0, 1, *$ 割り当てを行い，符号表を生成する．これにより，異なるカテゴリに属するサブカテゴリであっても，サブカテゴリ同士の特徴が類似したものは，同一のカテゴリとみなした符号表を生成することが可能となる．この 2 つの手順により，カテゴリ情報とサブカテゴリの特徴を考慮した符号表の生成を可能とする．その際生成される符号表の判別器数が少ないと分類精度の低下につながる．そのため，最後に 2 つの手順を繰り返すことにより符号表への追加を判別器数が所望の大きさとなるまで繰り返す．これにより，様々なサブカテゴリの作り方に対する二値分類器を用意した十分な判別器数を持つ 1 つの符号表を得る．

一方で，各繰り返しで生成されるサブカテゴリは異なるため，繰り返しにより得られる複数の符号表間において，同じ行番号であってもの同一のサブカテゴリではない．つまり，同じ行番号の符号語であっても同一のサブカテゴリに対するもではなく，符号表の追加は単純には行えない．そこで提案手法では，各行を学習データとする行数が総学習データ数となる符号表に変換を行う．これにより，同じ行番号にある符号語は同一の学習データに対する符号語となり，学習データごとに符号語の追加が行うことが可能となる．具体的な変換手順については，以下で示す．またサブカテゴリのグループを用いて符号表を生成する際には，あらかじめ K 行の符号表を準備しそれに基づき生成する．以下では全学習データ数を N, k 番目のカテゴリのサブカテゴリ数を S_k，あらかじめ準備する K 行の符号表を $\boldsymbol{H}$ とする．

Step1) サブカテゴリの代表ベクトルの選択　各カテゴリ c_k において，ランダムに S_k 個の学習データを選択し，それを各サブカテゴリの代表ベクトルとする．

Step2) サブカテゴリの生成　各カテゴリ c_k において，代表ベクトルと学習データとの距離を計算し，サブカテゴリの代表ベクトルとの距離が小さい学習データをサブカテゴリへと分割する．

Step3) サブカテゴリグループの代表ベクトルの選択　各カテゴリ c_k において，ランダムに一つのサブカテゴリを選択し，選択されたサブカテゴリの代表ベクトルを各サブカテゴリ遇グループの代表ベクトルとする．

Step4) サブカテゴリのグルーピング　サブカテゴリグループの代表ベクトルと各サブカテゴリの代表ベクトルとの距離を計算し，距離の近いグループへとサブカテゴリを分割する．

Step5) サブカテゴリグループに基づく符号表の生成　生成された K 個のサブカテゴリグループをあらかじめ準備した $\boldsymbol{H}$ の K 個のカテゴリとみなし，符号表を生成する．この際に，各サブカテゴリグループにおいてサブカテゴリを 1 つ選択し，そのサブカテゴリは判別に用いず符号語の要素は $*$ とする．上記の操作を，すべてのサブカテゴリが 1 回ずつ選択されるまで行う．

Step6) 符号表の結合　Step1 から Step5 を M 回繰り返す．各繰り返しで生成される符号表を，各行を学習データとした N 行の符号表に変換する．各学習データの符号語は属するサブカテゴリの符号語とする．変換後の M 個の符号表を並べてひとつの符号表とする．　□

3.3 l 多数決による分類基準

従来の類似度最大に基づく分類基準では，各カテゴリにおいて 1 つの符号語しか存在しないことを前提とした分類基準である．しかしながら，提案アルゴリズムにより生成される符号表は，各カテゴリにおいて複数の符号語を持つ．すなわち各カテゴリにおいて，新規データに対する判別器の出力と比較するテンプレートが複数存在することになる．そのため，類似度が最も高い符号語のみを考慮して分類を行うよりも，複数の符号語の類似度を考慮して分類を行う方がノイズに対して頑健になると考えられる．

そこで，提案する分類基準では，新規データに対する判別器出力と全符号語との類似度を従来と同様に計算し，その中から類似度の高い上位 l 個の符号語の属するカテゴリの多数決によってカテゴリを推定する．

4 実験・考察

提案した符号表生成アルゴリズムと分類基準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 2010 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 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，学習時間とし，それぞれ 5 回の実験の平均を用いた．二値判別器には RVM を用いた [2]．また，学習時間に関しては，各判別器の学習を並列処理した場合についても示す．提案手法のパラメータは，SUB_k はすべての k において 3, $M = 10$, $l = 3$ とし，$\boldsymbol{H}$ は一対他法とする．また比較手法は，Exhaustive 符号と一対他を用いる．各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	Exhasutive 符号	一対他法	提案手法
判別器数	255 (2 元)	9 (2 元)	270 (3 元)

表 3: 実験結果

符号表	正解率	計算時間 (秒)	
		通常	並列
Exhasutive 符号	0.763	58,822	330
一対他法	0.723	1,259	230
提案手法	0.776	15,711	113

表 3 より，正解率は提案手法が最も高くなり，計算時間は通常処理の場合は一対他が最も少なく，並列処理の場合は提案手法が最も少なくなった．

また，Exhaustive 符号と提案手法を比較すると，判別器数はほぼ同程度である一方で，計算時間，正解率ともに提案手法の方が良い値となった．計算時間に関しては，提案手法は 3 元符号表であるため，計算時間の低減が可能となったと考えられる．また，正解率に関しては，提案手法が一対他に基づく手法でありながら，高い精度となった．このことから，サブカテゴリを考慮した符号表生成と多数決分類法の有効性が確認された．

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，サブカテゴリを考慮した符号表生成とそれに伴う多数決分類基準の提案を行った．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．今後の課題としては，繰り返し処理やランダム性のない符号表生成アルゴリズムが挙げられる．

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