多元符号を用いたECOC法による多値分類に関する研究

1X14C094-3　西口 智之
指導教員　後藤 正幸

1　はじめに

近年，情報技術の発展により膨大な電子データが扱われるようになり，それらを自動で分類する技術の重要性が高まっている．分類問題では，カテゴリ情報が与えられたデータ集合を用いて識別規則を学習し，その規則に従ってカテゴリ情報が未知のデータをいずれかのカテゴリに分類する．ここで，3 つ以上のカテゴリが存在するものを多値分類問題と呼び，本研究ではこの問題を研究対象とする．多値分類問題に対する代表的なアプローチとして，単一の多値分類器を構成する手法と，複数の二値分類器を組合せる手法がある．

このうち，単一の多値分類器の代表的な手法として，Random Forest [1](以下，RF) がある．RF は，決定木を複数生成して統合し，説明変数と学習データをランダムに選択することにより，個々の木の多様性を高め，アンサンブルの効果を得る手法である．また，高い分類精度を示し，他の多値分類手法に比べ計算コストも少ないことから広く用いられている．

一方，複数の二値分類器を組合せる代表的な手法には Error Correcting Output Codes 法 [2](以下，ECOC 法) がある．ECOC 法は複数の二値分類器の出力を組合せることで多値分類問題に対応する手法であり，Support Vector Machine[3](以下，SVM) などの優れた二値分類器を活用しつつ，誤り訂正の機能を持たせることで分類精度の高い多値分類を実現する．

これらの双方のメリットを統合できれば，さらに高性能な多値分類器を構築できる可能性がある．ECOC 法の利点の 1 つとして複数のカテゴリを 1 つにまとめて識別境界を学習することで，分類問題がシンプルになり，個々の二値分類器で学習がし易い問題になっている点がある．そこで本研究では，ECOC 法の分類器構成方法を援用し，多値分類器である RF を複数組合せることで，従来よりも分類精度の高い分類器を構築する手法を提案する．従来の ECOC 法では多値分類問題を複数の二値分類問題に分解することで複数の二値分類器を構築していたが，本研究では多値分類問題を三値分類，四値分類などの部分問題に落とし込み，これらを統合することで，多値分類の精度向上を目指す．また，新聞記事データを用いた評価実験を行い，提案手法の有効性を示すとともに，結果に対する考察を行う．

2　分類問題

分類問題は，d 次元の特徴量ベクトル $\boldsymbol{x}_i \in \mathbb{R}^d$ にカテゴリ $y_i \in \mathcal{C} = \{C_1, \ldots, C_M\}$ の付与された i 番目の学習データ $(\boldsymbol{x}_i, y_i)$ を入力として学習を行い，カテゴリが未知である入力データ $\boldsymbol{x} \in \mathbb{R}^d$ に対応するカテゴリを推定する問題である．特徴量ベクトルの例として，文書データにおいては，単語頻度ベクトルを表す．M はカテゴリ数を表す．$M = 2$ の場合，二値分類問題と呼ばれ，$M \geq 3$ の場合，多値分類問題と呼ばれる．一般に，多値分類問題は分類カテゴリ数の増加に伴い問題が複雑化するため，分類精度を維持することが難しくなる．

2.1　Random Forest[1]

RF は，決定木を弱学習器として複数統合することで，汎化性能を向上させる手法である．本研究では RF を分類器として扱う．

RF の分類において新規データの所属カテゴリは，学習データにより学習された各弱学習器の出力の多数決によって推定される．そのため，カテゴリ間にデータ数の偏りが無い場合，RF は高い精度を示すことが知られている．

2.2　ECOC 法 [2]

ECOC 法は，符号理論で用いられる誤り訂正技術を多値分類問題に応用した手法であり，複数の二値分類器を構築し組合せることで，カテゴリ情報が未知の新規データの所属カテゴリを推定する．二値分類器には SVM 等が広く用いられている．ECOC 法は，SVM のような高性能な二値分類器を活用しようという発想から，これらの組合せにより多値分類問題に対応する手法として考案された．また，ECOC 法における二値分類器の構成は，符号表と呼ばれる {0,1} を要素とする二元符号表により表現される．二値分類器の個数を N とすると，符号表 $\boldsymbol{W}$ は $M \times N$ 行列で表される．符号表 $\boldsymbol{W}$ の列ベクトルは各二値分類器 f_n を表しており，要素が 1 のカテゴリ集合と要素が 0 のカテゴリ集合の二値分類を行う．行ベクトルは，各カテゴリ C_m の符号語であり，$\boldsymbol{w}_m$ と表す．ECOC 法における代表的な二値分類器の構成方法として one-vs-the rest 法がある．これは，1 つのカテゴリとそれ以外を分類する二値分類器を M 個組合せる手法である．以下の表 1 に one-vs-the rest 法の $M = 4$ における符号表を示す．

表 1．one-vs-the rest 法 ($M = 4$) における符号表

	f_1	f_2	f_3	f_4
$\boldsymbol{w}_1$	1	0	0	0
$\boldsymbol{w}_2$	0	1	0	0
$\boldsymbol{w}_3$	0	0	1	0
$\boldsymbol{w}_4$	0	0	0	1

二値分類器を組合せて最終的な分類カテゴリを決定する操作である復号の際は，各二値分類器の出力結果と符号語を比較し，最も距離の近いカテゴリに分類する．この距離を測る際には，硬判定の場合はハミング距離 [4] が用いられ，軟判定の場合には最尤推定法 [4] が用いられることが多い．

3　提案手法

3.1　提案への着想と概要

多値分類問題へ対応する手法として，RF などの多値分類器を用いた分類が一般的に有用であるが，カテゴリ数が多い場合，分類精度を維持することが難しいという問題がある．そこで本研究では，多値分類問題を複数の部分問題に落とし込むことで，分類精度を向上させる手法を提案する．これを実現するため，ECOC 法で用いられている様に，複数のカテゴリを 1 つのカテゴリにマージすることを考える．具体的には，全カテゴリの中から L 個のカテゴリを選び，それらをマージすることで，マージされたカテゴリを擬似的に 1 つのカテゴリとみなした部分問題へ帰着させ，部分問題に対応する多値分類器を構築する．この操作を考えられる全ての組合せについて行い，構築された分類器を組合せることで多値分類問題に対応する．これにより，各分類器の学習対象が部分問題となり，学習が容易になることに加え，マージされたカテゴリによる多様性が付加され，RF のアンサンブル効果が高まる可能性も期待される．

例えば $M = 4, L = 2$ とすると，4 つのカテゴリのうちの 2 つをマージしているため，擬似的なカテゴリ数は 3 となる．このため，RF を用いた個々の分類器が対象とするのはある 2 つのカテゴリとそれ以外を分類する三値分類問題という部分問題となり，このように生成された複数の三値分類器を組合せて分類する．これにより，4 つのカテゴリをそのまま四値分類問題として RF で分類した場合よりも，個々の RF が担当する分類問題の難易度が下がることが期待できる．このようにして得られる各分類器の分類精度が向上すると，未知

のデータに対する各分類器の出力値がより真の値に近づき，アンサンブルしたときの分類精度も向上する．

L 個マージする全ての多値分類器を組合せることで，分類器を構築する．このとき，必要となる多値分類器の数は $N = \binom{M}{L}$ となり，各多値分類器は $M - L + 1$ 値の分類器となる．以下に $L = 2, M = 4$ の場合の多元符号表を示す．

表 2．提案手法符号表 ($L = 2, M = 4$)

	f_1	f_2	f_3	f_4	f_5	f_6
$\boldsymbol{w}_1$	1	1	1	0	0	0
$\boldsymbol{w}_2$	2	0	0	1	1	0
$\boldsymbol{w}_3$	0	2	0	2	0	1
$\boldsymbol{w}_4$	0	0	2	0	2	2

上記の表 2 は，三値分類器の組合せ構成を表現している．f_1 は C_3 と C_4 を同一のカテゴリとみなした三値分類器となっており，同様に f_2 以降の分類器も 2 つのカテゴリをマージした分類器構成を表している．

入力データ $\boldsymbol{x}$ に対して，各分類器には RF を用いており，出力結果として各カテゴリ c に所属する確率 $P_n(c|\boldsymbol{x})$ が得られる．復号の際には，式 (1) で示す最尤推定法を用いて，カテゴリ $\hat{y}$ を推定する．

$$\hat{y} = \underset{c \in C}{\operatorname{argmax}} \sum_{n=1}^{N} \log P_n(c|\boldsymbol{x}_i) \tag{1}$$

4　評価実験

提案手法の有効性を検証するために，新聞記事データを用いた評価実験を行う．また，得られた結果に対して，分類精度を検証し考察する．

4.1　データ概要

本実験では新聞記事データをベンチマークデータとして使用した．以下の表 3 に，その概要を示す．

表 3．ベンチマークデータ (読売新聞，2015 年)

カテゴリ (数)	政治，経済，スポーツ，社会， 文化，生活，犯罪事件，科学 ($M = 8$)
文書の特徴 ベクトル (次元)	形態素解析による単語抽出 (40,067)
データセット	150 件/カテゴリ× 8 カテゴリ× 10 セット，

4.2　実験条件

本研究では，10 分割交差検証による評価を行う．また，評価指標としては，テストデータのうち正しいカテゴリに分類されたデータ数の割合を分類精度として用いる．

実験は $L = 1, 2, \ldots, 7$ の場合について行う．$L = 1$ は単一の RF，$L = 7$ は one-vs-the rest 法で従来手法となる．$L = 2, \ldots, 6$ が本研究における提案手法となる．$L = 7$ は RF が二値分類器になるため，SVM を用いた場合についても参考として比較する．

4.3　実験結果と考察

実験結果，及び各手法の分類器数を以下の表 4 に示す．

表 4．各手法実験結果

L	1	2	3	4
N	1	28	56	70
分類精度	0.825	0.841	0.853	0.860
L	5	6	7(RF)	7(SVM)
N	56	28	8	8
分類精度	**0.865**	**0.867**	0.862	0.866

表 4 より，単一の RF である $L = 1$ に比べて，全ての提案手法の精度が向上していることが分かる．また，RF を用いた one-vs-the rest 法 ($L = 7$) に比べ，$L = 5, 6$ のときに提案手法の方が優れている．加えて，$L = 6$ のケースは，SVM を用いた one-vs-the rest 法よりも優れている．一般に，二値分類器としての分類精度は RF に比べ，SVM の方が高いことが知られている．しかし，その SVM を用いた one-vs-the rest 法よりも $L = 6$ の方が優れており，カテゴリをまとめて部分問題を構成して学習することの効果を示している．

$L = 5$ の場合は 8 個あるカテゴリのうち，5 つのカテゴリをマージするため，個々の多値分類器は四値分類器となる．同様に $L = 6$ の場合は三値分類器となる．これらの結果より，三，四値分類器の RF を組合せることで，強力な二値分類器を組合せる $L = 7$ と同等かそれ以上の性能を示すことが分かる．

本研究での提案手法が，従来手法よりも高い分類精度が得られた理由として，八値分類問題を複数の部分問題に緩和した際に，より学習がし易い部分問題が構成されたことが考えられる．また，RF のアンサンブル効果が向上し，提案手法において各分類器の分類精度が向上したことも推測される．一方で，部分問題を構成する際に提案手法で過度に多くのカテゴリをマージすると学習データ量の偏りが大きくなる．例えば，$L = 7$ の時には二値分類器における 2 つのカテゴリは 1 対 7 の学習データ数の比率となり，データ量の少ないカテゴリの分類精度が低下してしまう．

このことから，マージするカテゴリ数は部分問題をどれだけ緩和するかということと，学習データ数のバランスのトレードオフの関係となっていると考えられる．本実験においては，個々の分類器の精度は $L = 6$，すなわち三値分類の時に最も高くなり，結果として，全体での分類性能が向上した．

5　考察

提案手法では複数のカテゴリをマージして，多値分類器を構築することで各分類器の分類性能の向上を実現した．本手法では単一の多値分類器では分類することが困難なデータを対象としている．すなわち，カテゴリ間の識別境界が学習しづらい問題である．そのため，容易に分類ができるデータに対しては分類性能の向上を期待できない．

しかし，現実問題には識別境界の学習が難しい問題が多く存在している．このような問題は文書分類問題だけでなく，画像データや手書き文字の認識問題にも介在しているため，提案手法は幅広い分野での運用が可能であると考えられる．

また，ECOC 法には二対他法や三対他法による符号表構成もあるため，それらと本提案との比較が必要である．

6　まとめと今後の課題

本研究では，多値分類問題を複数の部分問題に分解し，それらの部分問題に対応する RF の構築と，ECOC 法の概念に基づく，RF の出力のアンサンブルをする手法を提案した．これにより，RF と ECOC 法の持つ利点を併せ持つ手法の構築に成功した．また，提案手法を新聞記事データに適用することで，提案手法の有効性を示した．

本手法においては全カテゴリのマージを行うため，カテゴリ数の増加に伴い，計算コストが非常に大きくなる．そこでマージが必要となるカテゴリを予め識別し，そのカテゴリのみマージすることで計算量を削減する手法の提案などが今後の課題である．

参考文献

[1] Leo Breiman,“Random Forests,”*Machine Learning,* Vol.45, No.1, pp.5–31, 2001.

[2] T. G. Dietterich and G. Bakiri, “Solving Multiclass Learning Problems Via Error-Correcting Output Codes,” *Journal of Articial Intelligence Research,* vol.2, pp. 263–286, Jan.1995.

[3] C.Cortes and V. Vapnik, “Support-Vector Networks,”*Journal of Machine Learning Research,* vol.20, pp.273–297,1995.

[4] 平井有三,“はじめてのパターン認識”, 森北出版, 2012.