ECOC 法におけるリジェクトルールを用いた分類法に関する研究

情報数理応用研究　　5211C007-8　石橋想太郎
指導教員　後藤正幸

A New Classification Method Focusing on Reject Rule for ECOC Classification Systems

ISHIBASHI Sotaro

1　はじめに

近年，情報化社会の到来により，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る一方であり，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 Relevance Vector Machine (RVM) があげられ，優れた二値分類器として知られている [2]．本研究はこの RVM を用いて多値分類を行う方法を考える．RVM を多値分類問題に適用する場合，1 つの分類器で直接モデル化する方法，ならびに複数の二値分類器を組み合わせる方法の 2 つのアプローチが考えられている．前者は計算量の問題で実用的とは言えないため，従来から後者のアプローチの研究が多くなされている．その１つとして，符号理論の枠組みを導入した ECOC 法に基づく多値分類法 [3] がある．ECOC 法に基づく多値分類法は，対象となるカテゴリを二値分類器数の次元で構成される空間上の“ 符号語 ”に対応させ，二値分類器の出力結果からカテゴリを推定するものである．

一方で，上記で述べたような自動分類技術を利用する場合，現実には分類誤りが生じてしまうことがある．分類誤りが生じた場合，それを発見し，訂正しなければ正しい分類結果を得ることはできない．しかし，分類誤りを訂正するためのコストは一般に高いと言われている．そこで分類誤りを生じそうなデータに対してはその分類を行わず，より強力な分類器を利用するか専門家の判断に任せるというリジェクトと呼ばれる考え方がある [4]．

Simeone らはこのリジェクトの考えを ECOC 法に援用し，ECOC 法で組み合わせる１つ１つの分類器の出力に対しリジェクトルールを適用した．これにより信頼性の高い分類器の出力のみを用いた多値分類が可能となり，結果として分類精度の向上に成功している [5]．

ここで Simeone らが用いたリジェクトの判定基準（以下リジェクトルール）では，すべての分類器を同等に扱い，分類器間で同じ割合の出力をリジェクトするように棄却域が決定されている．ECOC 法はどの符号語にどのカテゴリを割り当てるかによって分類器間で分類誤りを生じる度合いが異なってくる．そのためすべての分類器を同等に扱う Simeone らの手法には改善の余地がある．また，誤り訂正符号には訂正能力と呼ばれるその符号固有の訂正可能な誤り個数の上限があるが，Simeone らの方法では用いた誤り訂正符号の訂正能力を超えた数の分類器の出力をリジェクトしてしまうことがある．これでは用いた符号の誤り訂正能力を利用することができず，ECOC 法の枠組みを十分に活かすことができない．

そこで本研究では,（１）分類器ごとの分類精度を考慮して，リジェクトする割合を変化させ，加えて（２）符号の誤り訂正能力を越えない様，分類器出力のリジェクト数に上限を設ける，と 2 つの視点から改良を加えたリジェクトルールを提案する．具体的には分類器の分類誤りが生じる度合いを予測精度によって定義し，これを分類器の“ 信頼度 ”とする．分類器の信頼度に合わせてリジェクトする割合を変化させ，より信頼度の高い分類器の出力のみを利用して ECOC 法の枠組みに適用する．また，リジェクトする出力の個数は用いる誤り訂正符号の訂正能力の数によって調整する．文書分類問題を対象とした実験を行うことで，従来手法より提案手法の分類精度が向上することを示す．

2　準備

2.1　多値分類問題

分類問題とはカテゴリラベルの付いた入力データを用いる学習を行い，新たに与えられた入力データ $\boldsymbol{x}$ に対応するカテゴリラベル $C \in \mathcal{C}$ をカテゴリ集合 $\mathcal{C}=\{C_1, C_2, \cdots, C_G\}$ から推定する問題のことである．ここで G はカテゴリ数を表し，多値分類問題とは $G \geq 3$ の場合を指す．多値分類のための手法としては大きく分けて 2 通りのアプローチが存在する．1 つはある 1 つの分類器で直接多値分類を行う方法であり，もう 1 つは強力な二値分類器を複数組み合わせて，その枠組みによって多値分類器を構成する方法である．本研究では後者を対象として研究を行う．

2.2　Relevance Vector Machine

RVM[2] は Tipping によって提案された手法で，回帰および分類問題を解くために提案された疎なカーネルベースのベイズ流学習手法である．優れた分類性能を持つ Support Vector Machine (SVM)[6] の特性の多くを引き継ぎながら確率モデルとして解釈できる点が最大の特徴である．

次に，二値 RVM の分類モデルを説明する．二値 RVM では M 個の学習文書セットを $\{\boldsymbol{x}'_k, t_k\}_{k=1}^{M}$ を用いて，入力データ $\boldsymbol{x}$ を，C_+ または C_- のどちらかのカテゴリに二値分類することを考える．ただし，$t_k \in \mathcal{C}$ である．ここで，$C_+, C_- \subset \mathcal{C}$, $C_+, C_- \neq \emptyset$, $C_+ \cap C_- = \emptyset$ とする．このとき $\boldsymbol{x}$ がカテゴリ $C \in C_+$ に分類される確率をロジスティック回帰関数を使って以下の式で表す．

$$p(C \in C_+ | \boldsymbol{x}) = \frac{1}{1 + \exp(-f_{\mathrm{RVM}}(\boldsymbol{x}))}, \tag{1}$$

$$f_{\mathrm{RVM}}(\boldsymbol{x}) = \sum_{k=1}^{M} w_k K(\boldsymbol{x}, \boldsymbol{x}'_k), \tag{2}$$

ただし，$w_k \sim N(0, \alpha_k^{-1})$ である．$K(.,.)$ はカーネル関数であり，入力された 2 つのデータ点を高次元空間上に写像し，内積を計算したものである．w_k は重み付けのパラ

メータであり，α の事後確率最大化により α_k^{-1} は推定されるが，その結果多くの w_k は 0 となる．w_k が 0 でないものを Relevance Vector(RV) と呼び，これらを用いて分類器 $f_{\mathrm{RVM}}(\boldsymbol{x})$ が構成される．RVM は高い汎化能力を持ち，出力が確率値で与えられる，カーネル関数が Mercer 条件[1]を満たす必要が無いなど多くの利点を持っている．この出力の確率値の値は，分類器がその分類に対してどの程度確信を持っているかの度合いを示す“ 確信度 ”を表していると考えることができる．

2.3 ECOC 法に基づく多値分類法

前述の通り，複数の二値分類器の組み合わせにより多値分類器を構成する方法には既に多くの有効な手法が提案されている．以下ではその代表的な従来手法として誤り訂正符号を用いる方法について述べる．誤り訂正符号 (ECOC) とは $\{0,1\}$ の二値で表現される情報系列に対して，より組織的に機械で処理しやすい形で冗長性を付加し，信頼性の向上を図る技術であり，多少雑音が混入しても元の情報に訂正することができる符号を指す．Dietterich らはこの手法を多値分類に援用し，多値分類問題を複数の二値分類問題に分解するための枠組みを与えた [3].

Dietterich らによる分類器構成法は Exhaustive 符号を用いるものであり，D を二値分類器の個数とし，分類器を $D = 2^{G-1} - 1$ 個作成する．Exhaustive 符号に基づく分類器構成法では，G 個のカテゴリに対し長さ D のベクトル $\boldsymbol{W}_{C_i}, (i = 1, 2, \ldots, G)$ を 1 対 1 対応させ，行列 $\boldsymbol{W} = [\boldsymbol{W}_{C_1}, \boldsymbol{W}_{C_2}, \ldots, \boldsymbol{W}_{C_G}]^{\mathrm{T}}$ を以下のように構成する．なお，$\boldsymbol{W}^{\mathrm{T}}$ は，$\boldsymbol{W}$ の転置を表す．

(1) $\boldsymbol{W}_{C_1}$ はすべて 1 で構成する．

(2) $i = 2, 3, \ldots, G$ において $\boldsymbol{W}_{C_i}$ は 2^{G-i} 個の連続する 0 と 1 を交互に D 個になるまで並べて構成する．

$G = 5, D = 15$ の場合の分類器構成を図 1 に示す．分類器の集合を $\boldsymbol{f} = \{f_1, f_2, \ldots, f_D\}$ とし，f_j が j 番目の分類器を表す．

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}
W_{C_1}	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
W_{C_2}	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
W_{C_3}	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1
W_{C_4}	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1
W_{C_5}	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0

図 1　$G = 5, D = 15$ の時の分類器構成

ここで $\boldsymbol{W}$ の各行は各カテゴリに対応する符号語となり，各列は各カテゴリをどのように分けるかを $\{0,1\}$ で表した分類器を意味する．例えば，図 1 の f_5 は，$\boldsymbol{W}$ の 5 列目に対応し，$C_+ = \{C_1, C_3\}$ と $C_- = \{C_2, C_4, C_5\}$ に分ける分類器を意味する．上記の Exhaustive 符号では，符号語をカテゴリ数である G 個作成して分類器を構成する．このとき符号長は分類器数 D となる．

分類法については，入力データ $\boldsymbol{x}$ に対する j 番目の分類器の $[0,1]$ を取る軟判定出力 R_j を用いて，カテゴリ C_i に対応する符号語 $\boldsymbol{W}_{C_i}$ の j 番目の値 W_{C_ij} が 0 ならば $1 - R_j$，1 ならば R_j を D 個の分類器の出力をかけあわせ，

$$\hat{C} = \arg\max_{C_i} \prod_{j=1}^{D} R_j^{W_{C_ij}} (1 - R_j)^{1 - W_{C_ij}}, \qquad (3)$$

とするカテゴリ $\hat{C}$ に分類する．これはカテゴリの事前確率が等確率のとき，C_i に対応した符号語の事後確率が最大となるカテゴリに分類することと等価である．

[1]通常カーネル関数が半正定値であることが必要である．

3 従来手法（Simeone らの手法 [5]）

一般的な分類問題において，現実には分類誤りが生じてしまう．分類誤りの訂正にはコストがかかるため，分類誤りを生じそうなデータに対しては分類を行わず，分類を棄却（リジェクト）することが考えられる．Simeone ら [5] はこのリジェクトの考えを ECOC 法に援用し，複数の分類器を組み合わせる ECOC 法において，分類誤りが生じる可能性のある分類器の出力は利用せず残りの正しく分類を行う可能性のある分類器の出力のみで ECOC 法の復号を行う方法を示した．これは信頼性の高い分類器の出力のみを用いて多値分類を行う方法であり，結果として分類精度が向上することが期待できる．以下ではこの Simeone らの手法について述べる．

3.1 リジェクトルールを用いた ECOC 法

2.2 節でも述べたように，RVM の出力は $[0,1]$ の軟判定値として与えられる．これは出力が 0.5 の近傍の場合，どちらのカテゴリであるかを判定する確信度が低いことを表している．入力データ $\boldsymbol{x}$ のカテゴリを決定する際，一般的な ECOC 法では符号の長さが D である場合 D 個全ての出力を利用する．しかしリジェクトルールを用いた場合では，分類器 f_j の出力 $R_j(\boldsymbol{x})$ を棄却域 $\mathcal{T}_j = [\tau_j^{lower}, \tau_j^{upper}]$ と比較し，棄却域内の出力は確信度が低いと判断しリジェクトとする．リジェクトされた分類器の出力は，その値を 0.5 とする．それにより，ECOC 法の枠組みに利用されずに消失していると考えることができ，残りの値を用いて軟判定復号を行う．以上をまとめると式 (4) のようになる．

$$R'_j(\boldsymbol{x}) = \begin{cases} 0.5, & \text{if} \quad R_j(\boldsymbol{x}) \in \mathcal{T}_j \\ R_j(\boldsymbol{x}), & \text{otherwise.} \end{cases} \qquad (4)$$

式 (4) から計算された値を用いて，2.3 節で述べた ECOC 法に基づく多値分類法を行う．

3.2 棄却域の決定方法

前節のリジェクトルールを用いる上で重要となるのは，リジェクトのための棄却域 $\mathcal{T}_j$ をどのように決定するかである．より多くの誤っている出力をリジェクトし，正しい出力をリジェクトしない棄却域を設定することが望ましい．従来手法では棄却域決定のための教師付きデータ（バリデーションデータ）を用意し，以下のステップでその棄却域 $\mathcal{T}_j$ を決定している．

Step1) 出力をリジェクトする割合 ρ を与える．ρ は全ての分類器において同一の値とする．

Step2) バリデーションデータに対して，それぞれの分類器の出力を得る．

Step3) バリデーションデータのうち，各分類器でリジェクトする出力の割合が ρ になるように棄却域 $\mathcal{T}_j$ の初期値を適当に定める．

Step4) リジェクトする割合を ρ で維持したまま，$\mathcal{T}_j$ の値を変化させ，$\mathcal{T}_j$ 内に含まれる正解データ数をカウントする．ここで正解データとは出力の値が $R_j(\boldsymbol{x}) > 0.5$ である正例と，$R_j(\boldsymbol{x}) < 0.5$ である負例である．

Step5) $\mathcal{T}_j$ 内に含まれる正解データ数が最も少なくなるような $\mathcal{T}_j$ を記憶し，その棄却域を分類に使用する．

前述の通り，この手法では ρ が全ての分類器において同一の値であるため，分類器ごとの信頼性の差を考慮できないという問題がある．

4 提案手法

Simeone らの手法ではリジェクトする割合 ρ をすべての分類器において同じとしている．しかし分類器ごとにその信頼性は異なり，信頼度が低い分類器の出力はより多くリジェクトし，信頼度が高い分類器の出力はできる限りリジェクトしない方が良い．また，従来手法を用いると 1 つの入力データに対するリジェクト数に上限が設けられていない．ECOC 法では，分類器の構成法に誤り訂正符号を用いることによって，その符号が持つ訂正の枠組みを援用している．しかし，リジェクトする個数が用いる誤り訂正符号の持つ訂正能力を超えてしまうと，分類器の構成に誤り訂正符号を用いた利点を失ってしまうという問題がある．

そこで本研究では，分類器の精度と誤り訂正符号の訂正能力を考慮した，新たなリジェクトルールを用いる分類方法を提案する．このリジェクトルールでは出力自体の確信度だけでなく，分類器自身の信頼度も考慮する．さらに，リジェクトする出力の個数は棄却域内のもののうちその値が 0.5 に近い上位 N 個までと上限を設け，この N は用いる誤り訂正符号の訂正能力の超えない個数に設定する．以上の方法により，分類器ごとの信頼度を考慮し，かつ ECOC 法の枠組みを有効に使うことのできる分類方法を提案する．

4.1 分類器の確信度と信頼度

RVM のような軟判定の二値分類器では，出力が $[0,1]$ の確率値として与えられる．この場合，0.5 近傍の出力はどちらのカテゴリであるか分類器が確信を持てていないことを表し，反対に 0 または 1 に近い場合はどちらのカテゴリか確信を持って分類できることを表す．そのため，分類器の出力の値はその分類の確信の度合い，すなわち“ 確信度 ”を表している．この確信度 $R_j(\boldsymbol{x})$ を指標として，分類器の信頼性を比較する．

具体的には，学習させた二値分類器の C_+ に含まれる正例の学習データに対する出力の平均によって，その分類器の信頼度 A_j を定義する．バリデーションデータとは別に用意された学習データの集合を $\mathcal{X}'$, 学習データ $\boldsymbol{x}'(\boldsymbol{x}' \in \mathcal{X}')$ のカテゴリを $C(\boldsymbol{x}')$，正例のカテゴリを持つ学習データ集合を $\bar{\mathcal{X}} = \{\boldsymbol{x}' | \boldsymbol{x}' \in \mathcal{X}', C(\boldsymbol{x}') \in C_+\}$ とし，分類器の信頼度 A_j を以下の式で計算する．

$$A_j = \frac{1}{|\bar{\mathcal{X}}|} \sum_{\boldsymbol{x}' \in \bar{\mathcal{X}}} R_j(\boldsymbol{x}'). \tag{5}$$

この“ 信頼度 ”は，その分類器がどれだけ正しく分類を行うかの度合いを表していると考えられる．信頼度が大きい程，分類器が正しく分類を行う可能性が高いことを示し，逆に信頼度が小さい程，正しく分類を行う可能性が低いことを示している．

4.2 誤り訂正符号とその訂正能力

誤り訂正符号の任意の二つの異なる符号語間のハミング距離の最小値を，この符号の最小ハミング距離または単に最小距離と呼ぶ．これは，符号の誤り訂正や検出の能力を決める重要なパラメータである．なお，符号の長さが n，情報記号数 k，最小距離が d_{min} の符号を (n, k, d_{min}) 符号とも書く．最小距離が d_{min} である符号を用いる場合は最大で，

$$t_0 = \left\lfloor \frac{d_{min} - 1}{2} \right\rfloor, \tag{6}$$

個の誤りまで訂正する復号が可能となる[2]．この t_0 を符号の誤り訂正能力と呼ぶ．Simeone らはこの誤り訂正能力を考慮せずに，リジェクトルールを ECOC 法に適用した．しかし訂正能力は ECOC 法の性能を決定する上で最も重要なパラメータであり，この点で Simeone らの手法には改善の余地がある．提案手法では，この訂正能力を維持できるように，リジェクト数を制限する．

[2] $\lfloor x \rfloor$ は床関数で x 以下の最大整数を示す．

4.3 信頼度を考慮したリジェクトルール

本研究では,（1）分類器ごとの分類誤りが生じる度合いに合わせて，リジェクトする割合を変化させる，加えて,（2）符号の誤り訂正能力を越えない様，分類器出力のリジェクト数に上限を設ける，2 つの視点で改良を加えたリジェクトルールを提案する．

具体的には分類器の信頼度 A_j を，テストデータに対する分類器の確信度 $R_j(\boldsymbol{x})$ に反映し，その計算結果を最終的な出力 $R'_j(\boldsymbol{x})$ とし，$R'_j(\boldsymbol{x})$ が棄却域内か否かを判断する．棄却域内である場合はその出力をリジェクトし，消失したものとみなして残りの出力で軟判定復号を行う．ここで，$R'_j(\boldsymbol{x})$ は式 (7) によって計算されるものとする．

$$R'_j(\boldsymbol{x}) = \frac{\{(2R_j(\boldsymbol{x}) - 1)A_j + 1\}}{2} \tag{7}$$

上記の方法に加えて，リジェクトする出力は棄却域内の $R'_j(\boldsymbol{x})$ のうち，その値が 0.5 に近い上位 N 個までと上限を設け，この N は用いる誤り訂正符号の訂正能力を超えない個数に設定する．これにより，誤り訂正符号の訂正能力を維持し，ECOC 法の枠組みの効果を活かして多値分類を行う．提案手法は以下のステップによって行われる．

Step1) 3.2 節の方法を用いて，棄却域 $\mathcal{T}_j$ を計算する．

Step2) 式 (5) により分類器ごとの信頼度 A_j を計算する．

Step3) テストデータに対する分類器ごとの出力 $R_j(\boldsymbol{x})$ を得る．

Step4) 式 (7) を計算し，確信度に加えて信頼度を考慮した出力 $R'_j(\boldsymbol{x})$ を得る．

Step5) $|R'_j(\boldsymbol{x}) - 0.5|$ を昇順にランキングし，j 番目の分類器の出力の持つ順位を S_j とする．棄却域 $\mathcal{T}_j$ と S_j から $R'_j(\boldsymbol{x})$ をリジェクトとするか否かを判断する．

$$R''_j(\boldsymbol{x}) = \begin{cases} 0.5, & \text{if } R'_j(\boldsymbol{x}) \in \mathcal{T}_j, S_j \leq N \\ R_j(\boldsymbol{x}), & \text{otherwise,} \end{cases} \tag{8}$$

この N を $N \leq t_0$ とすることで，符号の訂正能力を維持する．

Step6) R''_j を用いて 2.3 節で述べた ECOC 法に基づく多値分類法を行う．

5 実験による評価

5.1 新聞記事データを用いた分類実験とその方法

提案手法の有効性を検討するため，新聞記事データを用いて分類実験を行なった．実験には，読売新聞 2005 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) を使用した．すべての記事は 1 カテゴリのみに属し，カテゴリの重複はない．データから各カテゴリ 600 記事をランダムに選び，それを各カテゴリ学習データ 50 個，バリデーションデータ 50 個，テストデータ 500 個にランダムに分ける．学習データを用いて分類器を学習し，バリデーションデータで棄却域を定め，そ

れらを用いてテストデータの分類精度を評価する．以上の実験を 3 つのデータセットで行う．特徴量として，学習データに出現する全ての単語のうち，5 回以上出現する単語の単語頻度のみを使用する．ECOC 法において用いる誤り訂正符号としては，図 2 で表される (15,5,7)BCH 符号を用いた．BCH 符号は代表的な誤り訂正符号の 1 つであり，訂正能力の高い符号として知られている．$(15, 5, 7)$ 符号の誤り訂正可能数 t_0 は，$t_0 = 3$ である．ここではリジェクトする個数の上限を $N = 2$ としている．また，リジェクトする割合 ρ は，$[0, 0.3]$ の間で 0.05 刻みで変化させ，最も分類精度が高くなる値を用いた．

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}
W_{C_1}	1	0	1	0	0	1	1	0	1	1	1	0	0	0	0
W_{C_2}	0	1	0	1	0	0	1	1	0	1	1	1	0	0	0
W_{C_3}	0	0	1	0	1	0	0	1	1	0	1	1	1	0	0
W_{C_4}	0	0	0	1	0	1	0	0	1	1	0	1	1	1	0
W_{C_5}	0	0	0	0	1	0	1	0	0	1	1	0	1	1	1
W_{C_6}	1	1	1	1	0	1	0	1	1	0	0	1	0	0	0
W_{C_7}	1	0	0	0	1	1	1	1	0	1	0	1	1	0	0
W_{C_8}	1	0	1	1	0	0	1	0	0	0	1	1	1	1	0

図 2 用いた (15,5,7) BCH 符号

5.2 分類精度に関する比較

提案手法の有効性を検討するため，分類精度の比較評価を行なった．比較手法として二値判別器の簡単な組み合わせの方法である“ 1 vs the rest ”法，一般的な ECOC 法，従来手法であるリジェクトルールを用いた ECOC 法 (Simeone らの手法) を用いた．

表 1 はそれぞれの手法による分類の分類精度である．Simeone らの手法は一般的な ECOC 法より分類精度において優れている．このことから ECOC 法にリジェクトルールを適用することは，多値分類問題において効果があることがわかる．また，提案手法は分類精度において Simeone らの手法より優れている．このことから，分類器ごとの信頼度と符号の訂正能力を考慮する提案手法は有効であることがわかる．

表 1 分類精度 (読売，2005 年)

分類手法	1vsREST	ECOC 法	Simeone ら	提案手法
分類精度	0.619	0.637	0.643	0.651

5.3 学習データ数の変化による影響

学習データの数による提案手法の有効性の変化について調べるために，学習データ数を変化させた際の分類精度の評価を行った．ここでは，学習データの数を 50 件，100 件，150 件と変化させた．その他の実験条件は 5.1 節と同様である．

表 2 は，学習データ数を変化させた際の Simeone らの手法と提案手法の分類精度の変化である．学習データ数が増えるにつれて，Simeone らの手法と提案手法の分類精度の差は縮まっている．これは提案手法は分類器ごとの信頼度に差があるほど Simeone らの手法とは異なる出力をリジェクトするが，学習データ数が増加すると分類器の信頼度は全体的に高まり，分類器ごとの信頼度の差が明らかな形では現れず，結果として提案手法と Simeone らの手法のリジェクト結果が似たためだと考えられる．

反対に，学習データが少ないような状況や分類しづらいカテゴリが存在している状況においては，分類器ごとの信頼度の差が大きくなり提案手法を効果的に利用することができると考えられる．

表 2 学習データ数の変化による分類精度の変化

学習データ数	50 件	100 件	150 件
Simeone ら	0.619	0.637	0.643
提案手法	0.637	0.643	0.651

5.4 リジェクト上限数 N の寄与について

リジェクトする個数の上限は ECOC 法の枠組みにどのように影響するかを調べるために，リジェクトする個数の上限 N を変化させて実験を行った．ここでは，リジェクトする個数の上限 N を 1 個, 2 個, 3 個, ... と変化させた．その他の実験条件は 5.1 節と同様である．

表 3 はリジェクトする個数の上限 N を変化させたときの，提案手法の分類精度の変化である．N が用いた (15,5,7)BCH 符号の誤り訂正能力 $t_0 = 3$ を超えると，提案手法の分類精度が低下している．このことからリジェクトルールを ECOC 法に適用する場合は，誤り訂正能力を考慮してリジェクト数を制限する方がより効果的であることがわかる．

また，$N \leq t_0$ であっても N の値によって分類精度が異なるため，最適な N を決定する方法が別に必要になると考えられる．

表 3 N の変化による分類精度の変化

ECOC 法	$N = 1$	$N = 2$	$N = 3$	$N = 4$	$N = 5$
0.637	0.650	0.651	0.649	0.640	0.636

6 まとめと今後の課題

本研究では，リジェクトルールを用いた ECOC 法に関して，分類器ごとの信頼度と用いる誤り訂正符号の訂正能力を考慮した分類方法を提案し，実際の文書分類問題に適用することでその有効性を示した．

今後の課題は，文書分類問題以外の多値分類問題に提案手法を適用すること，リジェクトする個数の上限 N を最適に決定する方法を考案することである．

参考文献

[1] C. Silva and B. Ribeiro, “Scaling text classifcation with relevance vector machines, *Proc. IEEE International Conference on Systems, Man, and Cybernetics*, pp.4186–4191, Oct. 2006.

[2] M. E. Tipping, “Sparse bayesian learning and the relevance vector machine,” *Journal of Machine Learning Research*, pp.211–244, Jun. 2001.

[3] 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.

[4] C. Chow, “On optimum recognition error and reject tradeoff,” *IEEE Transactions on Information Theory*, pp.41–46, 1970.

[5] P. Simeone, C. Marrocco and F, Tortorella, “Design of reject rules for ECOC classification systems,” *Pattern Recognition*, pp.863–875, 2012.

[6] C. Cortes and V. Vapnik, “Support-vector net-works,” *Journal of Machine Learning Research*,vol.20, pp.273–297, Sep. 1995.