各カテゴリの学習データ数がアンバランスな場合の ECOC SVM による多値文書分類

1X10C113-0　安田直生
指導教員　後藤正幸

1　研究背景と目的

近年，Support Vector Machine（SVM）という二値判別器を用いた文書分類手法の有効性が示されている [1]．SVM を用いて多値判別問題を扱う場合には，複数の判別器を統合する．なかでも，頑健性の高い統合手法として，誤り訂正符号（Error-Correcting Output Code : ECOC）を利用し，2 つのカテゴリ部分集合を二値分類する判別器を複数用意して，学習後に統合する ECOC SVM が提案されている [2]．

一般に，SVM を含む二値判別器の学習では，二値判別を行う学習データ数がアンバランスでも，各カテゴリ集合の学習データの誤認識を平等に扱う．そのため，少数派の学習データにおける誤認識が，多数派のものと比較して割合的に高くなってしまい，新規文書データの分類にも悪影響を与えてしまう．ECOC SVM の場合，カテゴリ集合の構成により，各判別器で学習データ数のアンバランスの状況が異なる．この観点から，カテゴリ数のバランスがよい判別器のみを組み合わせることで精度を向上させる手法が提案されている [3]．しかし，各カテゴリのデータ数が異なるデータセットが用いられる場合，各二値判別器の学習データ数はやはりアンバランスになってしまう．また，各二値判別器を学習する際，両カテゴリ集合内の学習データはカテゴリごとに区別されない．そのため，各二値判別器の学習において，集合内でより学習データ数の多いカテゴリの影響を強く受け，それらの二値判別器を統合しても，学習データ数の少ないカテゴリの予測を適切に行うことができない．

本研究では，文書分類問題において各カテゴリの学習データ数がアンバランスな状況を仮定する．そして，SVM が二値判別を行う際，カテゴリ集合内とカテゴリ集合間のアンバランスがもたらす分類への悪影響を軽減することで，ECOC SVM の頑健性を保ちながら，学習データ数の少ないカテゴリにも精度よく分類を行うための手法を提案する．また，提案手法の有効性を，実際の新聞記事の分類問題に対する実験により検証する．

2　準備

2.1　多値判別問題

判別問題とは，カテゴリの情報を所持するデータを用いて学習を行い，新たに与えられたカテゴリが未知のデータ $\boldsymbol{x}$ に対し，その所属カテゴリ c_k（$k = 1, \ldots, K$）を推定する問題のことである．K はカテゴリ数を表す．多値判別問題は，$K \geq 3$ の場合の判別問題のことを指す．

2.2　Support Vector Machine

SVM は，分離超平面から最も近いデータまでの距離（マージン）を最大化するように二値判別を行う識別関数の学習手法である．マージン最大化によって汎化能力が高いという特徴があり,「高次元特徴空間」「文書ベクトルの点在性」といった文書分類問題の特性に起因する過学習という問題に対し有効とされている．

いま，入力ベクトルを $\boldsymbol{x}$ とする．このとき，各カテゴリ集合の学習データを分離する識別関数を，係数ベクトル $\boldsymbol{w}$，バイアス項 b を用いて，式 (1) で表す．

$$f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b. \tag{1}$$

識別関数 $f(\boldsymbol{x})$ を求めるために，各学習データ $\boldsymbol{x}_i$ に対して，カテゴリ集合ラベルを $t_i \in \{-1, +1\}$，カテゴリ集合ラベル側のマージンからの誤差距離（スラック変数）を ξ_i とし，次の最適化問題を解く．

$$\text{Mini. } \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} + C\sum_{i=1}^{N}\xi_i, \tag{2}$$

$$\text{s.t. } t_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) - 1 + \xi_i \geq 0, \tag{3}$$

$$\xi_i \geq 0. \tag{4}$$

ただし，C はスラック変数に対するペナルティパラメータ，N は学習データ数を表す．

3　従来手法 : ECOC SVM

ECOC SVM ではまず，全カテゴリを符号語の設定に従って 2 つのカテゴリ集合に分割する．そして，両カテゴリ集合に属する学習データを分類する二値判別器を学習し，それらを統合することで多値判別を行う．全ての分割の組み合わせを実現する場合，作成する二値判別器の個数 L は，

$$L = 2^{K-1} - 1, \tag{5}$$

で与えられ，判別器の構成は $K \times L$ 行列 $\boldsymbol{W}$ で表すことができる．行列 $\boldsymbol{W}$ の各行を L 次元ベクトル $\boldsymbol{W}_{c_k}$ とし，L 次元ベクトルで表現されるカテゴリ c_k（$k = 1, \ldots, K$）に対応する符号語とする．また，$\boldsymbol{W}$ の各列は各判別器における判別の仕方を意味し，カテゴリ集合ラベル $\{-1, +1\}$ に従いデータを二値判別することを意味する．全ての分割の組み合わせを用いる場合，$K = 4$ の判別器構成を図 1 に示す．

$$\begin{array}{c} \\ \boldsymbol{W}_{c_1} \\ \boldsymbol{W}_{c_2} \\ \boldsymbol{W}_{c_3} \\ \boldsymbol{W}_{c_4} \end{array} \begin{array}{c} \begin{array}{ccccccc} f_1 & f_2 & f_3 & f_4 & f_5 & f_6 & f_7 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 & 1 & 1 \\ -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ -1 & 1 & -1 & 1 & -1 & 1 & -1 \end{pmatrix} \end{array}$$

図 1. $K = 4$ の場合の判別器構成

$\boldsymbol{f} = \{f_1, \ldots, f_L\}$ は各二値判別器に対応し，例えば図 1 の f_1 は $\{c_1\}$ と $\{c_2, c_3, c_4\}$ を二値判別する判別器を表す．

各二値判別器を統合するためには，式 (6) の損失関数を用いる．

$$\epsilon_{kj}(\boldsymbol{x}) = \max(1 - g_{kj}D_j(\boldsymbol{x}), 0). \tag{6}$$

$g_{kj} \in \{-1, +1\}$ が j 番目の判別器でのカテゴリ c_k に対するカテゴリ集合ラベル，$D_j(\boldsymbol{x})$ がベクトル $\boldsymbol{x}$ を j 番目の判別器に入力した際の出力値を表し，$\epsilon_{kj}(\boldsymbol{x})$ が判別器 j から得られるカテゴリ c_k に対する誤りの度合いを表す．各判別器で $\epsilon_{kj}(\boldsymbol{x})$ を求め，それらの総和が最小となるカテゴリ $\hat{c}$ へと分類を行うことで，二値判別器の統合を行う．

$$\hat{c} = \arg\min_{c_k} \sum_{j=1}^{L} \epsilon_{kj}(\boldsymbol{x}). \tag{7}$$

4　提案手法

本研究では，各カテゴリの学習データ数がアンバランスな状況を仮定する．従来手法では，目的関数の最小化を考える際，各カテゴリ集合の学習データの誤認識を平等に扱う．そのため，少数派のカテゴリ集合の学習データに対する誤認識の割合が高くなり，少数派のカテゴリ集合の分類精度が低くなる．また，各二値判別器を学習する際に，両カテゴリ集合内の学習データはカテゴリごとに区別されない．そのため，二値判別を行う際，集合内でより学習データ数の多いカテゴリの影響を強く受けてしまい，ECOC SVM を用いて様々な二値判別器を統合しても，学習データ数の少ないカテゴリに対する予測を適切に行うことができなくなってしまう．

そこで提案手法では，各二値判別器において少数派のカテゴリ（またはカテゴリ集合）の学習データの誤認識に対する重みを強めることで，ECOC SVM における頑健性を保ちながら，カテゴリ集合内とカテゴリ集合間の 2 つのアンバランスによる影響をそれぞれ分けて解消することを考える．

4.1　Step1：カテゴリ集合内のアンバランスの解消

各二値判別器を学習する際，学習データセットにおける各カテゴリ c_k の各学習データを α_k 倍に複製する．

$$\alpha_k = \lfloor M/N_k + 0.5 \rfloor. \tag{8}$$

M は c_k が含まれるカテゴリ集合で最も多くの学習データをもつカテゴリの学習データ数，N_k は c_k の学習データ数を表す．学習データを複製することで，学習データ数の少ないカテゴリを考慮しつつ二値判別器を学習することを考える．

4.2　Step2：カテゴリ集合間のアンバランスの解消

各二値判別器を学習する際，少数派のカテゴリ集合の学習データの誤認識に対するペナルティパラメータ C を，多数派の β 倍となるように設定する ($\beta \geq 1$)．

$$\beta = N_{major}/N_{minor}. \tag{9}$$

N_{major}, N_{minor} はそれぞれ，多数派のカテゴリ集合の学習データ数，少数派のカテゴリ集合の学習データ数を表す．β を用いて，提案法の目的関数は次式で表すことができる．

$$\text{Mini. } \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} + C \sum_{t_i = t_{major}} \xi_i + \beta C \sum_{t_i = t_{minor}} \xi_i. \tag{10}$$

t_{major}，t_{minor} はそれぞれ多数派，少数派のカテゴリ集合ラベルを表す．異なるペナルティパラメータの設定を行うことで，少数派のカテゴリ集合に含まれるカテゴリを考慮しつつ二値判別器を学習することを考える．

5　実験

提案手法の有効性を検討するため，実際の新聞記事データを用いて分類実験を行い，評価を行った．

5.1　実験条件

実験には，毎日新聞 2005 年の 8 カテゴリ（社説・国際・経済・家庭・科学・芸能・スポーツ・社会）の記事を使用する．特徴量としては単語頻度を使い，文書頻度 5 以上の単語（名詞・動詞）によって特徴量空間を構成する．

表 1．各データセットでの各カテゴリのデータ数

	社説	国際	経済	家庭	科学	芸能	スポ	社会
$\mathscr{D}_1$	51	115	91	29	7	28	175	304
$\mathscr{D}_2$	63	111	93	47	30	46	157	253
$\mathscr{D}_3$	75	107	95	65	54	64	138	202
$\mathscr{D}_4$	87	104	98	82	77	82	119	151
$\mathscr{D}_5$	100	100	100	100	100	100	100	100

記事データ集合から，各カテゴリに対して表 1 に従った数の記事をランダムに選び学習データセットとする．ここで，$\mathscr{D}_1$ の構成が実際の記事数の比と等しく，$\mathscr{D}_2 \sim \mathscr{D}_5$ は評価のためにアンバランスの程度を変化させたデータセットである．また，各学習データセットに対し，学習データに選ばれなかった記事データから各カテゴリ同数の記事をランダムに選びテストデータセットとする．評価指標として，F 値（マクロ平均）を用いる．

$$\text{F 値} = \frac{1}{K}\sum_{k=1}^{K}\frac{2a_k}{A_k + T_k}. \tag{11}$$

a_k は c_k に分類され正解したテストデータ数，A_k，T_k はそれぞれ c_k に分類されたテストデータ数，c_k のテストデータ数を表す．比較手法として，ECOC SVM（従来手法），4.1 節の Step1 のみ（比較手法 1），4.2 節の Step2 のみ（比較手法 2）でデータセットのアンバランスを考慮した ECOC SVM を用いる．今回の実験では，図 1 のように全ての分割の組み合わせでの二値判別器の統合を行うこととする．

5.2　実験結果・考察

各データセットでの実験をそれぞれ 10 回ずつ実験し，F 値（マクロ平均）の平均を求めた結果を図 2 に示す．

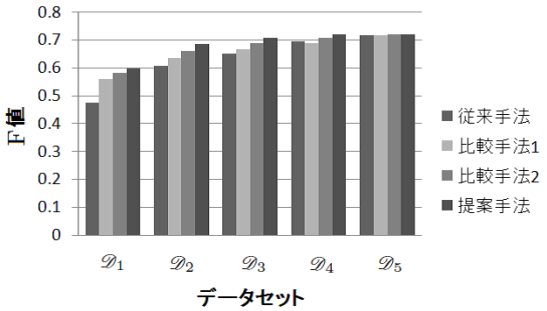

図 2．各データセットにおける実験結果

結果より，各手法と比較して，提案手法で F 値が向上し，提案手法の有効性を示すことができた．提案手法は各比較手法より F 値が高く，カテゴリ集合内とカテゴリ集合間の両方のアンバランスに対応することで，各カテゴリの学習データをより適切に扱うことができると考えられる．各カテゴリの学習データ数のバランスがよいデータセットでは提案手法による改善が薄れたが，これは提案手法によるアンバランスへの対応の効果が減少したためであると考えられる．

6　まとめと今後の課題

本研究では，各カテゴリの学習データ数がアンバランスな状況を仮定した ECOC SVM による多値判別手法について，多値判別のために組み合わせられる各二値判別器のカテゴリ集合内とカテゴリ集合間の双方のアンバランスに対応する手法を提案し，その有効性を示した．今後の課題として，3 元の符号語を用いた場合への拡張などが挙げられる．

参考文献

[1] JOACHIMS, Thorsten. “Text categorization with support vector machines: Learning with many relevant features,” Springer Berlin Heidelberg, 1998.

[2] 阿部重夫, “パターン認識のためのサポートベクトルマシン入門,” 森北出版株式会社, 2011.

[3] 小田井良輔，雲居玄道，三川健太，後藤正幸, “二値判別器の組み合わせによる RVM 多値文書分類手法に関する一考察,” 第 10 回情報科学技術フォーラム, pp. 425-428, 2011.