分類誤りに着目した RVM に基づく ECOC 判別器構成手法

1X11C030-0　尾崎 新之介
指導教員　後藤 正幸

1　研究背景と目的

本研究では，優れた二値判別器として知られる Relevance Vector Machine(RVM)[1] に対し，誤り訂正符号のアイディアを適用した ECOC 復号法によってテキストデータの多値判別を行う手法 [2] を対象とする．この手法は，効果的に構成された複数の RVM 二値判別器の出力を統合し，多値判別を行う手法であり，テキストデータを対象とした自動文書分類問題においても，その有効性が示されている．ECOC 復号法に基づく多値判別手法において,「個々の RVM が正例と負例のデータ数の偏りが少ない場合に性能が良いこと」,「各二値判別器の性能が高いという前提のもとでは，各行間の要素を列ごとに比べたときの異なる要素数であるハミング距離を大きく設定することで良い分類精度を示すこと」が知られている．小田井ら [3] は，BCH 符号を用いてこの 2 点を考慮した判別器構成を提案している．

しかし，文書カテゴリ間で単語の出現頻度が似通った単語も存在するため，誤分類を生じ易いカテゴリの傾向はカテゴリの統計的性質によって異なるが，小田井らの手法はこの点を考慮していない．そこで，実際に各判別器に学習させ，精度を測定した結果に基づき，適応的に判別器構成を変化させることで，精度を向上させることを考える．このような適応的な手法としては，金田ら [4]，Zhong ら [5] などの手法があるが，正例，負例のバランスを考慮した手法とはなっていない．本研究では，学習データのバランスを考慮しつつ，分類誤り率を低減させるように，適応的に判別器構成を生成することで，小田井らの手法の問題点の解決を試みる．具体的には，あるカテゴリのデータの誤分類先カテゴリの傾向に着目し，誤分類の傾向を持つカテゴリ間のハミング距離を適応的に大きくするよう判別器を追加することで，それらのカテゴリの分類精度を向上させる方法を提案する．特に，分類精度の低いカテゴリと最も誤分類データが多いカテゴリを分類し，正例，負例の比がほぼ等しい判別器の追加を行う点が特徴である．

2　準備

2.1　二値判別器と符号語

二値判別器の組み合わせによる多値判別手法では，二値判別器構成を符号表と呼ばれる 1，0 の二値からなる行列 $\boldsymbol{W}$ で表現する．行列 $\boldsymbol{W}$ は，判別器構成を表しており，p を二値判別器の個数，G をカテゴリ数とした場合，$G \times p$ 行列となる．行列 $\boldsymbol{W}$ の各行は符号語 $\boldsymbol{W}_{C_i}(i = 1, 2, \cdots, G)$ から構成され，各列が判別器 $\boldsymbol{f}_j(j = 1, 2, \cdots, p)$ に相当する．この判別器 $\boldsymbol{f}_j$ は，1 に対応するカテゴリ集合と 0 に対応するカテゴリ集合を二値判別し，$\boldsymbol{W}$ で構成された全判別器を用いて多値判別を行う．

2.2　Relevance Vector Machine

RVM[1] は特徴ベクトル間の類似度の性質を持つカーネル関数を用いた機械学習法であり，優れた二値分類器として知られる Support Vector Machine(SVM) の特性を引き継ぎつつ，確率モデルとして解釈できる点が特徴である．また，SVM の出力は識別結果であり，分類確率が 0.5 に近い値であっても 1，0 のどちらかを出力してしまうため，予測に対する事後確率を計算できなかったが，RVM を用いることで最大事後確率基準を用いることが出来るという利点がある．

2.3　“BCH half-vs-the rest”多値判別手法

BCH 符号とは，符号長 n，情報ビット数 k，誤り訂正可能ビット数 t をパラメータとして設定でき，一意に定まる符号である．この BCH 符号は，(n, k, t)BCH 符号と表される．(n, k, t)BCH 符号を符号語集合として用いた場合，符号長 n，符号語数 2^k, 最小ハミング距離が $2t + 1$ となる符号語集合が構築される．作成した最小ハミング距離が保障された符号語集合に対して，以下のアルゴリズムで符号語を選択することで，各判別器の 1，0 の比，各符号語間のハミング距離を考慮した判別器構成を得る．

Step0) $D = 2^k \geq G$ を満たす BCH 符号を用いて作成した，符号長 n の符号語を $\boldsymbol{d}_i = (d_{i1}, d_{i2}, \cdots, d_{in})$，符号語集合を $\Delta = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_D\}$ とする．

Step1) 符号語集合の重みベクトル $\boldsymbol{Q} = (q_1, q_2, \cdots, q_n)$ を式 (1) で計算する．

$$q_s = \sum_{h=1}^{D} d_{hs} \tag{1}$$

Step2) 重みベクトルから目標ベクトル $\boldsymbol{V} = (v_1, v_2, \cdots, v_n)$ を式 (2) で作成する．

$$v_s = \begin{cases} 1, & q_s \geq \max\limits_i q_i - 1 \\ 0, & q_s < \max\limits_i q_i - 1 \end{cases} \tag{2}$$

Step3) 目標ベクトルと各符号語で，ハミング距離が最も小さい符号語を削除する．

Step4) $D = G$ になるまで Step1 から繰り返す．

Step5) G 個の符号語を各カテゴリに対応させ，同一の分け方の判別器を削除する．

また，入力 $\boldsymbol{x}$ に対する各判別器の RVM による出力を $\boldsymbol{R} = (R_{C_1}, R_{C_2}, \cdots, R_{C_G})$ とすると，ある入力 $\boldsymbol{x}$ に対して，カテゴリ C_i の事後確率 Y_{C_i} は式 (6) で表され，

$$Y_{C_i} = \prod_{k=1}^{p} R_k^{W_{C_i k}} (1 - R_k)^{1 - W_{C_i k}} \tag{3}$$

Y_{C_i} が最大となるカテゴリ C_i に判別する．

3　提案手法

3.1　小田井らの手法の問題点

新聞記事などの分類において,「犯罪事件」といったカテゴリでは特徴的な単語の頻度が多く，正しく分類されやすい．しかし,「生活」や「文化」といった日常的な話題がテーマとなりやすいカテゴリ間など，データの特徴が似通っているため誤分類されやすいカテゴリも存在する．小田井らの手法では各カテゴリの符号語に差はなく，全てのカテゴリを同等に扱っているため，このような誤分類の傾向を考慮できないという問題点がある．

3.2　適応的な判別器構成手法の拡張

実際に各判別器に学習させた結果に基づき，適応的に分類を行う手法のアイディアを利用し，小田井らの手法 [3] の問題点の解決を試みる．適応的に符号表を構成する方法として，金田らの手法 [4] や Zhong ら [5] の手法がある．金田らの手法 [4] では，判別器の出力が誤り訂正できないデータと学習データの経験分布の差を考慮した判別器構成を適応的に生成する．また，Zhong らの手法 [5] では，学習データから各二値判別器が分割する 2 つの部分空間の特徴を考慮した部分空

間を適応的に構成する．本研究では，小田井らの手法と同様に学習データの正例，負例のバランスと，誤分類の傾向を考慮しつつ，適応的に判別器を追加し，判別器構成を得る．

3.3 カテゴリペアの選択

提案手法では，まず“BCH half-vs-the rest”多値判別手法により学習データを学習し，検出用データを分類する．その分類結果から，設定した閾値 α よりも分類精度が低いカテゴリとそのカテゴリの最も多い誤分類先のカテゴリをカテゴリペアとして抽出する．ここで，閾値 α よりも分類精度が低いカテゴリ全てを抽出カテゴリとし，各抽出カテゴリに対し，最も多い誤分類先カテゴリを着目カテゴリと定義する．

3.4 分類誤りを考慮した判別器の追加

抽出カテゴリと着目カテゴリの符号語間のハミング距離を大きくするように判別器を追加する．以下のアルゴリズムにより，小田井らの手法をもとに分類誤りの傾向，各符号語間のハミング距離，各判別器の 1，0 の比の全てを考慮し，適応的に判別器を追加することで分類精度向上を図る．

Step0) 学習データを用いて各二値判別器を学習し，“BCH half-vs-the rest”多値判別手法を適用し，検出用データを分類する．

Step1) 検出用データに対する分類精度が閾値 α よりも低いカテゴリ (抽出カテゴリ) とそのカテゴリで最も誤分類データ数の多いカテゴリ (着目カテゴリ) をカテゴリペアとして抽出する．

Step2) 全てのカテゴリペア内で 1，0 の要素が異なるような組み合わせが存在する場合はそれらの要素を固定し，Step4 へ．存在しない場合は Step3 へ．

Step3) 最も精度の高い抽出カテゴリとその着目カテゴリのカテゴリペアを削除し，Step2 へ戻る．

Step4) Step2 で与えられたカテゴリペア以外の要素について 1，0 の比が等しくなる判別器の組み合わせを全て作成する．これらの判別器の重複を削除し，残りを判別器群とする．

Step5) Step4 で作成した判別器群を“BCH half-vs-the rest”多値判別手法の判別器構成に追加する．

{ 抽出カテゴリ，着目カテゴリ } が $\{C_3, C_6\}$，$\{C_5, C_6\}$，$\{C_6, C_3\}$ となる場合の，追加される判別器群を以下に示す．

$$
\begin{array}{c} \\ \boldsymbol{W}_{C_1} \\ \boldsymbol{W}_{C_2} \\ \boldsymbol{W}_{C_3} \\ \boldsymbol{W}_{C_4} \\ \boldsymbol{W}_{C_5} \\ \boldsymbol{W}_{C_6} \\ \boldsymbol{W}_{C_7} \\ \boldsymbol{W}_{C_8} \end{array}
\begin{array}{c}
\begin{array}{cccccc} f_{17} & f_{18} & f_{19} & f_{20} & f_{21} & f_{22} \end{array} \\
\begin{pmatrix}
1 & 1 & 1 & 1 & 1 & 1 \\
1 & 0 & 0 & 1 & 1 & 1 \\
1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 1 \\
1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 1 & 1 \\
0 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 1 & 1 & 0 & 0
\end{pmatrix}
\end{array}
$$

図 1．提案手法で追加される判別器群の例 ($G = 8$)

4 実験

4.1 実験条件

読売新聞 2005 年のデータから 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，各カテゴリの学習データを 50 件，検出用データを 100 件，テストデータを 100 件ランダムに抽出したものを 1 つのデータセットとして扱い，3 つのデータセットにおける実験を行う．比較手法として従来の“BCH half-vs-the rest”多値判別手法 (比較手法 1)，追加した判別器の有効性を示すため“BCH half-vs-the rest”多値判別手法に加え，提案手法と判別器数が等しくなるようにランダムに構成された判別器を追加した手法 (比較手法 2) を用いる．提案手法で用いる閾値 α は，Step0 で検出用データを分類したときの分類精度とし，BCH 符号のパラメータは $(n, k, t) = (31, 6, 7)$ とする．比較手法 1，比較手法 2 では，学習データで学習を行い，テストデータを分類する．提案手法では，学習データ，検出用データを用いて判別器構成を生成し，テストデータを分類する．また，比較手法 1 と提案手法では 3 つのデータセットにおける分類精度の平均を，比較手法 2 では各データセットにおいて判別器のランダム追加を 5 回行い，計 15 回分の分類精度の平均を比較する．評価指標に用いる分類精度は，正解カテゴリに分類されたテストデータ数の割合とする．

4.2 結果と考察

表 1．分類精度

	比較手法 1	比較手法 2	提案手法
分類精度	0.626	0.676	0.691

表 2．カテゴリペア

抽出カテゴリ	経済	スポーツ	文化	生活
着目カテゴリ	生活	文化	スポーツ	文化

表 3．抽出カテゴリ別の分類精度

	経済	スポーツ	文化	生活
比較手法 1	0.520	0.460	0.450	0.420
比較手法 2	0.572	0.458	0.490	0.422
提案手法	0.620	0.450	0.530	0.470

表 1 より，提案手法が最も高い精度を示すことが確認できた．比較手法 1 に比べ，比較手法 2 が高い精度を示したのは，判別器数増加のためだと考えられる．表 3 より，「経済」，「文化」，「生活」のカテゴリでは提案手法が高い精度を示したが，「スポーツ」のカテゴリでは低い精度を示した．「スポーツ」の着目カテゴリは「文化」であったが，「政治」，「生活」においても同程度の誤分類データがあることが確認できた．提案手法では，最も誤分類データの多いカテゴリである，「スポーツ」のカテゴリとのハミング距離を大きく設定したが，「政治」，「生活」とのハミング距離も大きく設定すべきであり，追加した判別器が有効でなかったと考えられる．

5 まとめと今後の課題

本研究では，RVM 二値判別器の正例，負例のバランスを考慮した方法を出発点としつつ，分類誤りを考慮した適応的な判別器構成手法を提案し，分類精度向上を確認した．精度の低いカテゴリとその分類誤り先カテゴリのハミング距離を大きくする判別器を追加することで，誤分類されるデータの多かったカテゴリの精度が向上し，全体の精度が改善したと考えられる．今後の課題として，各カテゴリで学習データ数に偏りがある場合やカテゴリペアの選択基準をカテゴリによって変えた場合での有効性の確認などが挙げられる．

参考文献

[1] M. E. Tipping, “Sparse Bayesian Learning and the Relevance Vector Machine,” *Journal of Machine Learning Research*, pp. 211–244, Jun. 2001.

[2] 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.

[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.

[4] 金田有二，上田修功，“誤り訂正符号を用いた多重カテゴリ分類,” 電子情報通信学会技術研究報告，PRMU, パターン認識・メディア理解 102(379), pp.13-18, 2002.

[5] Guoqiang Zhong, Mohamed Cheriet, “Adaptive Error-Correcting Output Codes,” *Proc. IJCAI ’13 Proceedings of the Twenty-Third international joint conference on Artificial Intelligence* pp.1932-1938 , 2012.