事後確率最大判別法に基づく RVM 多値文書分類手法の提案

1G06H028-1　小田井良輔
指導教員　後藤正幸

1　研究背景

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，中でも高精度の文書自動分類技術が必要とされている．

最近，文書分類の分野ではカーネル学習を用いた方法の性能が非常に高いと報告されている．その代表的な手法として，Relevance Vector Machine (RVM) があげられる [1]．しかし RVM は優れた二値判別器として知られているが，多値判別の方法がまだ確立されていない．そこで本研究では RVM が確率モデルであることを利用して，二値判別器の冗長構成と事後確率の計算による多値分類の手法を提案する．提案手法の有効性を示すため，実際の新聞記事の分類問題に適用し，分類精度が向上することを検証する．

2　準備

2.1　多値判別問題

判別問題とはカテゴリラベルの付いた入力を使って学習を行い，新たに与えられた入力 $\boldsymbol{x}$ のみからこれに対応するカテゴリラベル $C \in \{C_1, C_2, \cdots, C_G\}$ を推定する問題のことである．G はカテゴリ数を表し，多値判別問題とは $G > 2$ の場合の判別問題のことを指す．

多値判別の手法としては，大きく分けて 2 通りのアプローチが存在する．1 つは多値判別問題を 1 つの判別器で直接モデル化するものである．もう 1 つの手法は複数の二値判別器の組み合わせで多値判別器を構成するものである．

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．分類性能の良い Support Vector Machine (SVM) の持つ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 RVM の分類モデルを説明する．入力ベクトルを $\boldsymbol{x}$，カテゴリラベルを $t \in \{0, 1\}$，N 個のトレーニング文書セットを $\{\boldsymbol{x}_n, t_n\}_{n=1}^{N}$ とする．このとき $t = 1$ となる確率をロジスティック回帰を使って以下の式で表す．

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

$$f_{RVM}(\boldsymbol{x}) = \sum_{i=1}^{N} w_i K(\boldsymbol{x}, \boldsymbol{x}_i). \tag{2}$$

$K(.,.)$ は入力された 2 つのデータ点を高次元空間上に写像し，その内積を表すカーネル関数である．w_i は重み付けのパラメータであり，平均 0，分散 α_i^{-1} の正規分布に従う確率変数である．事後確率最大化により α_i^{-1} は推定されるが，その結果ほとんどの w_i が 0 となる．w_i が 0 でないものを Relevance Vector (RV) と呼び，これらを用いて決定関数 $f_{RVM}(\boldsymbol{x})$ を構成する．

RVM は高い汎化能力を持ち，出力値が確率値である，カーネル関数が Mercer 条件を満たす必要が無いなど多くの利点を持っている．RVM の大きな問題は学習に時間がかかってしまうことである．しかし RVM は RV の数が少なくなりコンパクトなモデルが得られるので，実際の応用で重要となる判別時間は一般的に短いという特徴を持つ．

3　従来手法

3.1　従来の RVM 多値判別手法

RVM 多値判別手法を 1 つの判別器で直接モデル化する方法は計算コストが非常に大きく実用的ではない．カテゴリ数 $G > 2$ クラスの問題では，学習にかかる計算量が 2 クラス RVM の G^3 倍になるといわれている．一方，複数の二値判別器の組み合わせで多値判別器を構成する方法は多くの有効な手法が提案されているため後者の枠組みを取り扱う．

RVM 多値判別手法では従来，$i = 1, 2, \cdots, G$ のそれぞれに対して，カテゴリ C_i とそれ以外のカテゴリに分ける判別器を作る．入力 $\boldsymbol{x}$ に対する各々の判別器の出力を $\boldsymbol{R} = (R_{C_1}, R_{C_2}, \cdots, R_{C_G})$ とすると，

$$\hat{C} = \arg\max_{C_i} R_{C_i}, \tag{3}$$

とするカテゴリ $\hat{C}$ に判別する．

3.2　ECOC 復号法に基づく多値判別法

誤り訂正符号 (ECOC) は情報系列にパリティ系列と呼ばれる冗長な情報を付加し，符号語として扱うことにより，情報を伝達する際に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．

Dietterich と Bakiri は ECOC に基づき，多値判別問題を複数の二値判別問題に分解するための枠組みを与えた [3]．p を二値判別器の個数とした場合，G 個のカテゴリラベルをそれぞれ p 次元ベクトルの符号語 $\boldsymbol{W}_{C_i}$ に 1 対 1 対応させる．その多値判別法では，符号語 $\boldsymbol{W}_{C_i}$ と入力 $\boldsymbol{x}$ に対する p 個の二値判別器の $\{0, 1\}$ の硬判定出力のハミング距離を H_{C_i} とし，

$$\hat{C} = \arg\min_{C_i} H_{C_i}, \tag{4}$$

とするカテゴリ $\hat{C}$ に判別する．

4　提案手法

4.1　背景

従来の RVM 多値判別手法の問題点として，

- 1 つの判別器の精度が悪いだけで，全体の分類性能が悪くなってしまう，
- 1 対多判別器では，多カテゴリの学習データ数に比べ，1 カテゴリの学習データ数が少なくなってしまうため，良い判別器が作れなくなってしまう場合が多い，

という 2 つの問題点が挙げられる．一方，ECOC 復号法に基づく多値判別法では，各判別器の信頼性の差異を全く考慮せずに $\{0,1\}$ の硬判定で判別するために有効に働かない場合がある．

これらの 3 つの問題を改善するため，学習データ数の偏りを少なくしつつ冗長な判別器を作る．そして，各カテゴリラベルを符号語とし，$\{0,1\}$ の硬判定のハミング距離で評価するのではなく，RVM が確率モデルである特性を用いて軟判定である事後確率最大判別法によって，カテゴリを判別する手法を提案する．

4.2 判別器構成法

従来手法の判別器 G 個に加えて，$\lceil G/2 \rceil$ 個と $\lfloor G/2 \rfloor$ 個に分けるような判別器を全ての組み合わせで作る．ただし，$\lceil x \rceil$ は x 以上の最小の整数，$\lfloor x \rfloor$ は x 以下の最大の整数である．ただし，カテゴリ数が偶数の場合，2 つの組 $\{A, B\}$ に分けたものと $\{B, A\}$ に分けたものは判別器としては同値であるために，新たに作る判別器の個数 $g(G \geqq 4)$ は，

$$g = \begin{cases} \frac{G!}{(G/2)!^2 \times 2}, & G \text{ が偶数の場合}, \\ \frac{(G+1)!}{((G+1)/2)!^2 \times 2}, & G \text{ が奇数の場合}, \end{cases} \tag{5}$$

である．例えば $G = 4$ の場合は，カテゴリ数を 2 個と 2 個に分けるような組み合わせは $g = 3$ であるので，判別器数は $G + g = 7$ 個である．

4.3 判別方法

ある入力 $\boldsymbol{x}$ に対して，カテゴリ C_i の符号語 $\boldsymbol{W}_{C_i}$ の k 番目の値 $W_{C_i k}$ が 0 ならば $1 - R_k$，1 ならば R_k を p 個の判別器の出力をかけあわせたものを Y_{C_i} とし，以下の式で表す．

$$Y_{C_i} = \prod_{k=1}^{p} R_k^{W_{C_i k}} (1 - R_k)^{1 - W_{C_i k}}. \tag{6}$$

このとき，

$$\hat{C} = \arg\max_{C_i} Y_{C_i}, \tag{7}$$

とするカテゴリ $\hat{C}$ に判別する．

5 実験方法

提案手法の有効性を検討するため，新聞記事の実データを用いて分類実験を行い，分類精度の評価を行なった．

5.1 実験条件

実験には，毎日新聞 2005 年の 4 カテゴリ (社会・スポーツ・芸能・経済) の記事を使用する．すべての記事は 1 カテゴリだけに属し，カテゴリの重複はない．データから各カテゴリ 250 記事ずつの合計 1000 文書をランダムに選び，それを学習データ各カテゴリ 150 個，テストデータ 100 個にランダムに 2 回分ける．これを 4 回繰り返し，計 8 回の実験結果で評価を行なう．提案手法は判別器数を (5) 式より 3 個増やし計 7 個とした．特徴量としては単語頻度を使い，文書頻度 50 以上の単語頻度を特徴量として使用する．カーネル関数は (8) 式で表される線形カーネルを用い，d は自然数であり，$d = 1$ とした．

$$K(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}\boldsymbol{y} + 1)^d. \tag{8}$$

比較手法として，

- 3.1 節で記した 4 個の判別器で判別する従来手法 1，
- 提案手法と同じく計 7 個の判別器でその出力が 0.5 以上なら 1，それ以外なら 0 として，3.2 節で記したハミング距離で判別する従来手法 2，

を用いた．

5.2 実験結果

従来手法 1・従来手法 2・提案手法の分類精度の実験結果を図 1 に示す．結果より，提案手法が従来手法 1 と従来手法 2 の両方よりも分類精度で勝っており，提案手法の有効性を示すことができた．

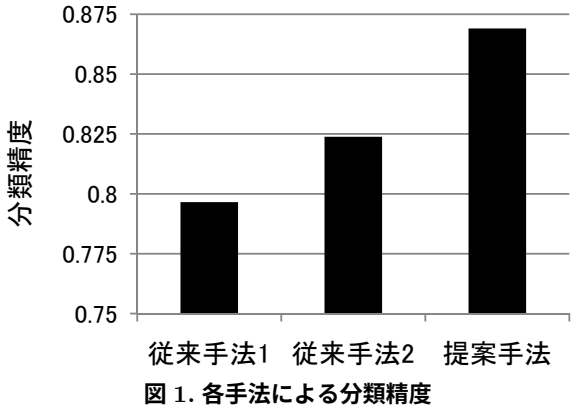

図 1. 各手法による分類精度

5.3 考察

提案手法は今回は文書分類に適用したが，判別器を増やして各カテゴリを符号語とし，事後確率を計算する方法なので，他の様々な RVM を使った多値分類問題にも応用することができると考えられる．

6 まとめと今後の課題

本研究では RVM による多値判別手法について，冗長な判別器を作り，RVM の確率モデルを使って分類する手法を提案し，その有効性を示した．

今後の課題は，判別方法の計算で判別器の分類精度で出力に重み付けをする方法を検討する必要がある．さらに，文書が複数のカテゴリに属することを考慮する場合にどのように対応するかについても今後検討が必要である．

参考文献

[1] C.Silva and B.Ribeiro, “Scaling Text Classification with Relevance Vector Machines,” *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, 2001.

[3] 大山賀己, 竹之内高志, 石井信, “ＥＣＯＣ復号法に基づく階層的多値判別法,” 電子情報通信学会, 電子情報通信学会誌, vol. 107, pp. 337–342, 2008.