F-015

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
A Study of multivalued document classification based on combination of binary RVM classifiers

小田井良輔* 雲居玄道† 三川健太* 後藤正幸‡
Ryosuke Odai Gendo Kumoi Kenta Mikawa Masayuki Goto

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．また何らかの情報を記録しようとする場合，現時点では，記述形式の柔軟性や計算機に蓄積する際のデータ量という観点から，文書データの形式をとることが最も一般的な手法であろう．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．

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 [1]．その代表的な手法として，Relevance Vector Machine (RVM) があげられ，優れた二値判別器として知られている [2]．しかし，多値判別の問題に適用する際，1 つの判別器で直接モデル化する方法は可能であるが，計算量の問題で実用的とは言えない．この問題を回避するため，”1-vs-the rest”多値判別手法と呼ばれる方法が知られている [3]．これは，1 つのカテゴリを識別する二値判別器をカテゴリ数だけ用意する方法である．しかしながら，各カテゴリを識別する二値判別器を用意する方法は，簡単な方法である反面，学習データの偏りによって，判別器の性能悪化を招くなどの問題がある [4]．

このような二値判別器の組み合わせで多値判別を実現する方法のとして，ECOC 復号法に基づく多値判別法 [4] や BT モデルを用いた方法 [5] が提案されている．これらは，多値判別問題を複数の二値判別問題に分解する枠組みを与えたものである [3]．本研究ではより広く用いられている前者の枠組みの立場で議論する．

ECOC 復号法に基づく多値判別法は硬判定で判別するために，確率値で出力を与える RVM のようなモデルの長所を生かすことができない．そこで筆者らは RVM が確率モデルであることを利用して，複数の二値判別器の組み合わせと事後確率の計算による多値分類の手法を提案している [8]．提案法では，各二値判別器が分類するカテゴリの学習データ数バランスを考慮した構成法を与えており，分類精度の面で優れていることが示されている．しかし，多値判別に対する二値判別器の代表的な構成法である，ECOC 復号法に対する優位性については検討の余地があった．そこで本研究では，実際の文書分類問題に対し，硬判定分類器を対象とした ECOC に基づく多値判別法との性能比較を行い，提案手法の有効性を検証する．

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 $\mathcal{C} \in \{C_1, C_2, \cdot\cdot, C_i, \cdot\cdot, C_G\}$ を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，多値判別問題とは $G \geq 3$ の場合を指す．

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り，もう一方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．前述の通り，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) [9] 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 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 $\mathcal{C} \in \{C_1, C_2\}$，$N$ 個のトレーニング文書セットを $\{\boldsymbol{x}_n, t_n \in \{C_1, C_2\}\}_{n=1}^{N}$ とする．このとき $C = C_1$ とな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．

$$p(C = C_1|\boldsymbol{x}) = \frac{1}{1 + \exp(-f_{RVM}(\boldsymbol{x}))}, \quad (1)$$

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

ただし，$w_i \sim N(0, \alpha_i^{-1})$ 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_i 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_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，出力が確率値で与えられる，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．一方，RVM は SVM と比較して学習により多くの時間を要するという問題点がある．実際，M 個の基底関数を持つモデルを用いると，$M \times M$ 行列の逆行列を計算するため，RVM の学習には $O(M^3)$ の時間がかかる [6]．

多値判別問題については，G 個の線形モデルを組み合わせる確率的な方法を用いる．α_i^{-1} は 2 クラスの場合と同じように計算する．この方法は一貫性があるという点では有利ではあるが，学習にかかる計算量が 2 クラス RVM の G^3 倍となってしまう点が不利である [6]．

3 従来手法

前述の通り，RVM を用いて多値判別手法を 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

*早稲田大学大学院創造理工学研究科
†早稲田大学理工学術院総合研究所
‡早稲田大学理工学術院

ではない．一方，複数の二値判別器を組み合わせ，多値判別器を構成する方法は既に多くの有効な手法が提案されている．本節では，これらの従来手法について述べる．

3.1 ”1-vs-the rest”多値判別手法

二値判別器を複数用いて，多値判別を行う方法の代表的手法は，”1-vs-the rest”多値判別手法である [3].

”1-vs-the rest”多値判別手法では，全てのカテゴリ $i = 1, 2, \cdots, G$ に対して判別対象カテゴリ C_i とそれ以外のカテゴリに分ける”1-vs-the rest”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符号を指す．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 [4].

p を二値判別器の個数,G をカテゴリラベル数とした場合，判別器構成を表す $G \times p$ 行列を $\boldsymbol{W}$，行列 $\boldsymbol{W}$ の各行を p 次元ベクトル $\boldsymbol{W}_{C_i}, (i = 1, 2, \cdots, G)$ とし，カテゴリ C_i の符号語とし，1 つのカテゴリ C_i に対応させる．符号語 $\boldsymbol{W}_{C_i}$ は $\{0, 1\}$ で構成され，$\boldsymbol{W}$ の各列は判別器の分け方を意味し，それぞれ対応する $\{0, 1\}$ を判別する．

Dietterich と Bakiri による判別器構成法は Exhaustive Codes を用いるものである [4]．判別器を $p = 2^{G-1} - 1$ 個作成する．Exhaustive Codes に基づく判別器構成法としては，$\boldsymbol{W}_{C_1}$ は全て 1 で構成する．$\boldsymbol{W}_{C_2}$ は 2^{G-2} 個の 0 に続き $2^{G-2} - 1$ 個の 1 で構成する．$\boldsymbol{W}_{C_3}$ は 2^{G-3} 個の 0，2^{G-3} 個の 1，2^{G-3} 個の 0 に続き $2^{G-3} - 1$ 個の 1 で構成し，$\boldsymbol{W}_{C_i}$ は 2^{G-i} 個の 0 と 1 を交互に並べて構成する．$G = 5$ の場合の判別器構成を図 2 に与える．

	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．$G = 5$ の時の判別器構成

$\boldsymbol{f} = \{f_1, f_2, \cdots, f_p\}$ はそれぞれ判別器であり，例えば，図 1 の f_5 は $\{C_1, C_3\}$ と $\{C_2, C_4, C_5\}$ を判別する．

判別方法は，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．

この手法は p 個の二値判別器の精度が同等のとき，性能が良いとされている．

4 提案手法

4.1 問題設定と背景

本研究では，各カテゴリに所属する学習データの数が全て等しく，データが各カテゴリから出力される確率が全て等しいという問題設定とする．このとき，”1-vs-the rest”多値判別手法の問題点として，

- 1 つでも判別器の性能が低いと，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，良い判別器を構成できない可能性が高い，

という 2 つの問題点が挙げられる．

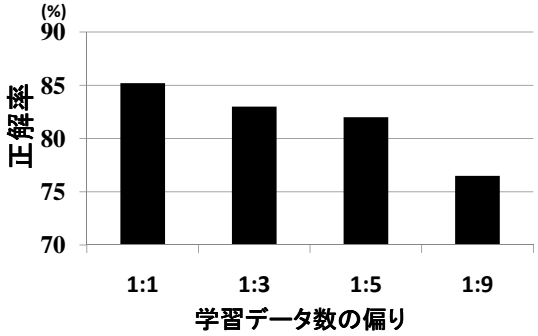

図 2．学習データの偏りによる分類精度

図 2 は RVM を用いた多値判別手法について，学習データの偏りを加味した際の正解率である．実験データは毎日新聞 2000 年のデータを用いた．実験は，学習データ数を 300 個に固定し,「正解カテゴリに属する学習データ」と「正解カテゴリに属さない学習データ」の比をそれぞれ 1 : 1,1 : 3,1 : 5 と 1 : 9 にすることで，それぞれの正解率を算出した．図 1 から，カテゴリに所属するデータと所属しないデータの数が等しいとき，データの偏りが大きくなるにつれて正解率が低下していく傾向が見てとれる．

一方，ECOC 復号法に基づく多値判別手法では，各判別器の信頼性の差異を全く考慮せずに，RVM の確率値出力を $\{0, 1\}$ の硬判定を用いて判別するために有効に働かない場合がある [8]．さらに学習データの偏りを考慮していない判別器構成のため，”1-vs-the rest”多値判別手法と同じく良い判別器を構成できない可能性が高い．

以上の問題を改善するため，本研究では学習データ数の偏りを少なくした冗長な多値判別器構成ならびに，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

提案手法は，判別器構成として，図 1 より，1 : 1 が 1 番正解率が良いので，カテゴリを $\lceil G/2 \rceil$ 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 は $G \geq 4$ のとき，

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

となる．

このような構成法で作った方法を提案法とし，”1-vs-the rest”多値判別手法に冗長な判別器を加えたことによる性能の向上と，新たに作る偏りの小さい判別器による性能の向上を確かめるために，比較手法として，新たに作った g 個の判別器に，”1-vs-the rest”多値判別手法の G 個を加えたものを考える．

以下にカテゴリ数 $G = 6$ の時の判別器構成の例を示す．

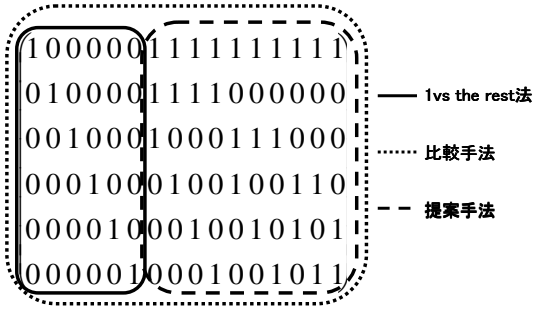

図 3．$G = 6$ の時の判別器構成

$G = 6$ のとき，従来手法である判別器数は”1-vs-the rest”多値判別手法で $p = G = 6$，比較手法では $p = G + g = 16$，提案手法では $p = g = 10$ である．

ここで，

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 番目の判別器構成の値 $W_{C_i k}$ 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 実験による評価

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 2 つの方法の分類実験を行い，分類精度の評価を行なった．

5.1 実験方法

実験には，毎日新聞 2000 年の 8 カテゴリ (国際 · 経済 · スポーツ · 社会 · 芸能 · 家庭 · 総合 · 文化) の記事と読売新聞 2000 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，同様の実験を行った．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 550 記事をランダムに選び，それを各カテゴリ学習データ 500 個，テストデータ 50 個にランダムに分ける．特徴量として，学習データに出現する全ての単語の単語頻度を使用する．カーネル関数は (8) 式で表される線形カーネルを用い，$d = 1$ とした．

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

また，学習データ数，カテゴリ数を変えた際の性能変化について実験を行う．前者は，データとして毎日新聞 4 カテゴリ (国際 · 経済 · スポーツ · 社会)，読売新聞 4 カテゴリ (政治 · 経済 · スポーツ · 社会) の記事を使用し，学習データの数を各カテゴリ，100 記事から 500 記事まで 100 記事ずつ増やしていき，5 パターンで行うものであり，後者は，毎日新聞はカテゴリ数 G は $G = 4$(国際 · 経済 · スポーツ · 社会)・$G = 6$($G = 4$ のカテゴリ + 芸能 · 家庭)・$G = 8$(全カテゴリ) の 3 パターン，読売新聞は $G = 4$(政治 · 経済 · スポーツ · 社会)・$G = 6$($G = 4$ のカテゴリ + · 文化 · 生活)・$G = 8$(全カテゴリ) の 3 パターン，計 6 パターンで実験を行う．従来手法として，3.1 節で記した G 個の判別器で判別する”1-vs-the rest”多値判別手法と 3.2 節で記した $2^{G-1} - 1$ 個の判別器で判別する ECOC 法を用いた．

5.2 学習データ数変化実験結果

毎日新聞・読売新聞の記事データそれぞれに対し，1-vs-the rest 法・ECOC 法・比較手法・提案手法の正解率の実験結果を図 3，図 4 に示す．学習データ数が増加するにつれ，全体的に正解率も向上している．学習データ数 100 件においては，どの手法でも明確な差はない．

図 4 より比較手法は，学習データ数が 300 件のときを除き，従来手法より優れている．

提案手法は全てのパターンにおいて，判別器数が最も少ないにも関わらず，他の全ての手法よりも同等以上の精度であることがわかる．これらから提案手法の有効性を示すことができた．

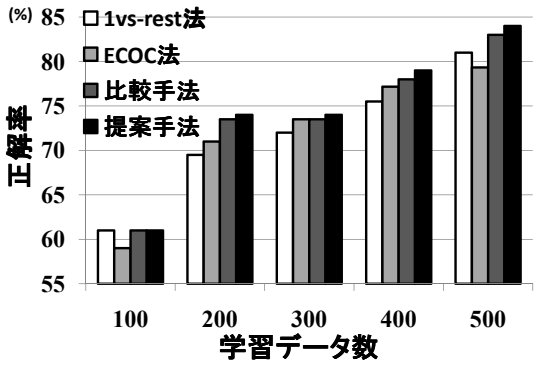

図 4．学習データ数と分類精度 (毎日新聞)

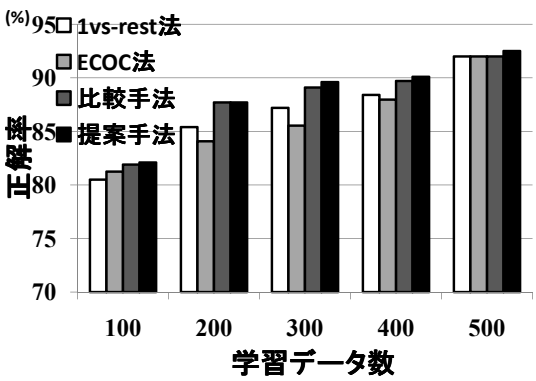

図 5．学習データ数と分類精度 (読売新聞)

5.3 カテゴリ数変化実験結果

毎日新聞・読売新聞の 1-vs-the rest 法，ECOC 法，比較手法，提案手法の正解率の実験結果を図 6，図 7 に示す．提案手法は全てのカテゴリにおいて，比較手法よりも優れている．全てのカテゴリにおいて，提案手法が従来手法より優れているため，提案手法の有効性を示すことができた．

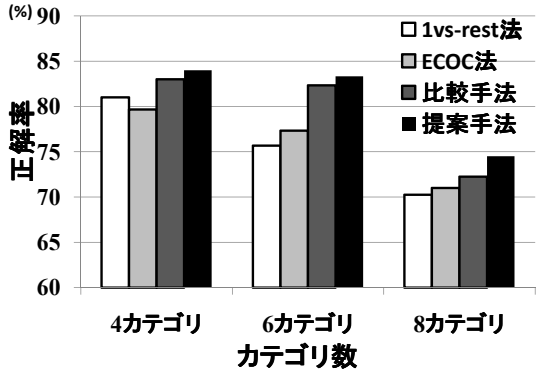

図 6．カテゴリ数の違いと分類精度 (毎日新聞)

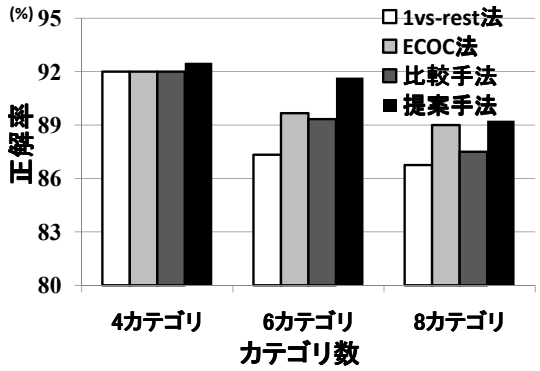

図 7．カテゴリ数の違いと分類精度 (読売新聞)

5.4 考察

本研究で実施した 2 種類の実験において，比較手法や ECOC 法よりも判別器数が少ないにも関わらず，提案手法の分類精度が最も高かった．分類精度の差については以下の理由が考えられる．

- 提案手法以外の判別器構成では，カテゴリ数が大きくなるにつれて，学習データ数に偏りが発生している．このため，各判別器の精度が下がり，比較手法，1-vs-the rest 法・ECOC 法で分類精度が下がってしまったと考えられる．
- ECOC 法は判別器数が最も多いにも関わらず，精度は良い結果でなかった．これは RVM の軟判定ではなく，硬判定であるハミング距離を用いて判別したため分類精度が上がらなかったと考えられる．

また，提案手法で新たに作った判別器の中でも，判別器の精度にはばらつきが存在するために，判別器の精度が高いものだけを使って判別すれば，さらに分類精度が向上すると考えられる．

今回の実験では，最大 8 カテゴリの実験としたため，冗長に作成した判別器数は最大 35 個であった．一方，RVM を 1 つの判別器で直接モデル化する場合，計算量は 2 クラス判別器の 3 乗 (512) となるので，提案手法を用いることで学習時間が削減できたと考えられる．しかし，例えば 20 カテゴリの実験を行う場合，今回の提案手法では，92378 個の判別器を作成する必要がある．これは，学習に要する時間が膨大となってしまい現実的な数字ではない．カテゴリ数が 15 以上となった場合，提案手法よりも 1 つの判別器で直接モデル化する場合の方が学習時間が短いため，提案手法は G が少ないときにより適した手法であると考えられる．

6 まとめと今後の課題

本研究では，実際の文書分類問題に対し，硬判定分類器を対象とした ECOC に基づく多値判別法との性能比較を行い，提案手法の有効性を検証した．

今後の課題は，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，他の様々な多値分類問題にも容易に応用する必要があると考えられる．また，学習量の面において，G が大きくなると提案手法は有効ではない．よって，判別器の精度が良いものだけを選択し判別器を構成する方法を検討する必要がある．

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, Jun. 2001.

[3] 池田思朗, “2 クラス判別器の組み合せによる多クラス判別統計モデルとパラメータ推定,” 統計数理研究所, 特集「統計的機械学習」, vol.2, pp. 157–166, 2010.

[4] 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

[5] T.Hastie and R.Tibshirani, “Classification by pairwise coupling,” *The Annals of Statistics*, vol.26, pp. 451–471, Apr.1998

[6] C.M. ビショップ, “パターン認識と機械学習 下,” シュプリンガー・ジャパン, pp.56–67, 2008.

[7] 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.

[8] 小田井良輔, 谷口祐樹, 雲居玄道, 後藤正幸, “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,” 経営情報学会　全国研究発表大会要旨集, Vol. 2010f , pp. 61–64, Nov. 2010.

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