Vol.2012-MPS-87 No.16
2012/3/2

文章の識別モデルを状態とする HMM によるテキストの段落分割

但　馬　康　宏[†1]

テキストセグメンテーションに対する新たな分類モデルを提案する．一般に HMM によるテキストセグメンテーションは，状態を段落に対応させ，各状態においては，その段落にふさわしい単語の出力確率を学習することにより文書をモデル化する．本研究では，HMM の状態を段落に対応させる点はそのままとし，各状態において 1 つの文章を受け入れる識別モデルを構成することにより文書全体を認識するモデルを提案する．すなわち本手法における HMM の各状態は，その状態が表す段落にふさわしい文章を高確率で受け入れ，ふさわしくない文章を低い確率で受け入れる事により段落を表現するモデルとなる．評価実験として，ウェブニュース記事の分類問題を行い，従来手法よりも性能が向上することを確かめた．

A text segmentation method via HMM and discriminative models

Yasuhiro Tajima[†1]

We propose a new method for a text segmentation problem via HMM to recognize a text. HMM has been applied to a text segmentation problem in previous studies. Then, each state represents a topic and a word is the output symbol on the state. In our method, HMM has states by which topics are represented as same as previous studies, but each state has a discriminative model to recognize a sentence. We evaluate this method by the news text segmentation problem. Then, we have confirmed that our method performs beter than the previous method.

†1 岡山県立大学 情報システム工学科
Department of Systems Engineering, Okayama Prefectural University

1. はじめに

テキストセグメンテーションの問題に対して，HMM を用いた文書のモデル化手法を新たに示す．テキストデータを段落や章，話題など意味のある分割位置で区切ることをテキストセグメンテーションもしくは，段落分割と呼ぶ．この問題に関して，従来 2 通りの研究方針が広く知られている．第一の方法は，TextTiling[1] などで広く知られている変化点を抽出する手法である．まず，テキストデータに対し一定の範囲のテキスト窓を切り取り，その窓内のテキストを特徴付ける特徴量を算出する．テキスト窓をテキストの先頭から末尾まで動かしてゆき，特徴量の変化が大きい位置が分割位置であるとする手法である．例えば特徴量として，あるテキスト窓内に現れる単語の種類とその出現数をベクトルにしたものを考えると，ひとつの窓と隣接する窓との間には，2 つのベクトル間のなす角を類似度とみなすことができる．窓を動かしてゆき，類似度が大きく変動する位置が，大きく話題の転換する位置だとみなせ，分割位置の候補となる．この手法では，どの程度の変動を分割位置とするかという閾値問題など設定すべきパラメータが性能に大きな影響を与える．事前の学習にあたる部分がない点が特徴である．

第二に HMM を用いた分割手法である[4]．一般的には，単語を 1 つの出力記号とし，HMM の各状態が 1 つの段落や話題を表すものとする．音声認識の分野では音素の抽出などに広く使われており，時系列データの処理での性能の高さがよく知られている．事前に学習データを用いてパラメータを設定することが多く，Baum-Welch などのアルゴリズムが知られている．この手法は，いくつかの発展形があり，状態に到着した時点で出力する記号を確率変数の長さをもった記号列とし，テキストセグメンテーションに適した改良を行う研究[3] や，出力記号と前状態から現在状態を決定する HMM (MEMM) への改良[2] などがある．いずれの研究においても，HMM の各状態は段落を表し，出力記号が一単語であるので，段落に対する単語ユニグラムによる言語モデルを構築している．

本研究では，HMM の各状態を 1 つの文章識別器としてテキストを認識する手法を提案する．この手法は，各状態が段落や話題を表す点は従来手法と同じだが，各状態では，1 つの文章を確率的に識別するものとする．すなわち，分割対象のテキストについて，1 文ごとに各状態での受け入れ確率が求められるものとし，テキスト全体において最も受け入れ確率が高い状態遷移系列を求め，互いに違う状態への遷移が段落の切れ目であるとする手法である．状態遷移確率は一般の HMM と同じ扱いができ，それぞれの文に対する受け入れ確率の和が，それぞれの状態で 1 となるならば，本手法においても Baum-Welch アルゴリズム

を利用することができる．

評価実験として，複数のウェブニュースが連なったテキストファイルに対してニュースの記事ごとへの分割を行った．その結果，本手法により従来手法よりも高い性能を得ることができ，特にランダムに話題が移り変わるようなテキストデータに対しては，大きな性能向上となることが確認できた．

2. 提 案 手 法

2.1 HMM による段落分割とその改善

実数の集合を R とする．離散型隠れマルコフモデル (HMM) を状態の有限集合 Q，出力記号の集合 B，状態間の遷移確率 $a : Q \times Q \to R$，各状態における出力確率 $b : Q \times B \to R$ にて定義する．任意の $i \in Q$ について，$a(i, \cdot)$ および $b(i, \cdot)$ は確率分布である．初期状態確率分布を $i \in Q$ について $a(0, i)$ と表す．

テキスト t は単語の列 $w_1 w_2 \cdots w_n$ であり，扱うすべてのテキストに出現するすべての単語の集合を W と表す．一般に HMM を用いたテキスト分割は，学習データであるテキスト集合 T を用いて単語の出力モデルである HMM を構成し，分割対象のテキストに対し最適な状態遷移系列を求め，その状態の移り変わりが話題の移り変わりであると見なし，分割位置を決定する．

HMM の各パラメータの推定には，EM アルゴリズムである Baum-Welch アルゴリズムがよく知られている．この学習アルゴリズムは，教師なし学習アルゴリズムでありサンプルデータの集合から直接 HMM の各パラメータを推定することができる．すなわち，テキストに現れる話題を 1 つの状態とし，その話題を述べる場合に出現しやすい単語の分布を出力記号の分布としてモデル化する手法である．

この手法の発展として，テキストの段落ごとに話題のラベルを付けた学習データから，話題のラベルを出力記号とする HMM を構成し，段落分割を行う手法を提案した[5]．この手法では，学習データを 1 文ごとに分割し，文とラベルとの関係から 1 文に対してどのラベルを割り当てるべきかを決定する分類器を構成する．さらに，学習データであるテキストを 1 文ごとに 1 つのラベルが付いたラベルの記号列に変換し，そのラベルの記号列を出力する HMM を構成する．分割対象のテキストに対しては，分類器を用いて 1 文ごとにラベルを推定し，ラベルの列を作成する．次に作成したラベルの列を生成する最適な状態遷移系列を学習により構成した HMM を用いて推定し，各文がどの話題であるかを決定し，段落分割を行う．この手法では，学習に話題のラベルが付いた学習データが必要だが，前記の単語を出力記号とする HMM による分割よりも高性能であることが確認できた．

2.2 本研究における手法

本研究では，以下の視点にもとづき HMM を用いた段落分割手法を改善する．

- 段落の切れ目は必ず文の終わりであり，文の途中で区切られることはない．
- 複数の単語の組み合わせで特徴的な用語となる場合がある．

以上の点から，一文を分割できない範囲と見ることにより，分割性能の向上が期待できる．

一般に，複数の単語を含む範囲を取り扱うには n-gram を出力記号とする HMM とすることが考えられる[6]．しかし，n-gram の出現確率は，単語 1 つの出現率 p の場合に比べ p^n となるため，より多くの学習データが必要であり，学習時間も増加する．本研究では 1 つの文章を出力記号とし，各状態における出力確率は，別に準備した文章の識別器を用いて決定する． すなわち，HMM における状態遷移確率は従来と同じく，状態 i, j について

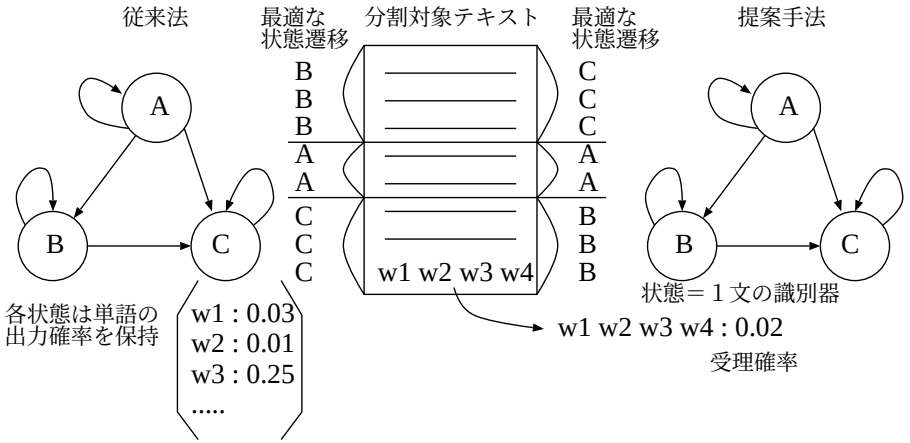

図 1 本手法における HMM と従来手法との比較

$a(i, j)$ と表され，各状態は話題を表すが，出力記号は 1 つの文章となる．図 1 に本研究における HMM と従来手法による HMM との違いを示す．

本研究における文章の識別器は，ナイーブベイズを用いた識別器とした．これは，HMM 全体の学習において，学習アルゴリズムが正しく収束することを保証することができたため採用した．ナイーブベイズ以外の手法でも，HMM の状態遷移および初期確率の学習が矛

盾なく行え，各状態が対応する段落に対する文章の識別器として機能するならば利用可能である．

具体的には，本研究における HMM は状態の有限集合を Q，状態間の遷移確率を $a : Q \times Q \to R$ とし，初期状態確率分布をすべての $i \in Q$ について $a(0,i)$ と表すまでは，従来法と同じである．したがって，任意の $i \in Q$ について $a(i,\cdot)$ は確率分布である．文章を識別する識別器は，以下のように構成される．各状態 $i \in Q$ は，学習データに現れたすべての単語 $w \in W$ について，1 つの実数値 $p_i(w)$ を保持する．文章 t を n 単語からなるとして，$t = w_1w_2w_3 \cdots w_n$ と表す．状態 i における文章 $t = w_1w_2w_3 \cdots w_n$ に対する識別確率 $p_i(t)$ は，

$$p_i(t) = p_i(w_1) \cdot p_i(w_2) \cdot p_i(w_3) \cdots p_i(w_n)$$

として求める．この確率が，一般の HMM における出力確率に対応する．各状態で保持する単語の受け入れ確率 $p_i(w)$ は，確率分布である必要はなく，与えられた文章に対して識別確率を求められればよい．しかし制約が何もない場合，学習アルゴリズムによっては，すべての単語の受け入れ確率が 1 となるような場合も考えられる．その場合，状態遷移をせず，すべての文章を受け入れるのが最適状態遷移系列となってしまう．したがって，1 つの状態における総和

$$\sum_{w \in W} p_i(w)$$

は一定の値となるように正規化して利用する必要がある．本研究においても上記の正規化を行い Baum-Welch アルゴリズムを適用し，各パラメータを決定した． 図 2 に本研究における HMM が 1 つのテキストデータを認識する過程を示す．

以上より，本研究における HMM に対する Baum-Welch アルゴリズムは，以下のようになる．

(1) 状態遷移確率と初期確率 $\forall i, j, a(i,j)$，各状態における単語の受け入れ確率 $\forall i \in Q, p_i(w)$ について，初期値を設定する．
(2) 前向き，後向きアルゴリズムを用いて，n 番目のサンプルの l 行目への到達確率 $\gamma(n,l,i)$ を求める．
(3) それぞれのパラメータを更新する．

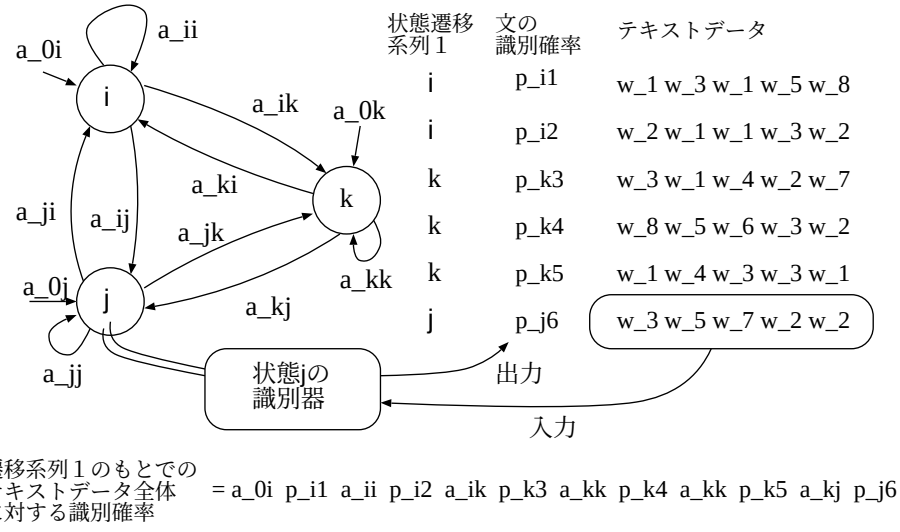

図 2 本手法における文書の出力過程

3. 評 価 実 験

3.1 実験データ

評価実験として，ウェブのニュース記事をつなげたものを 1 つのテキストとし，このテキストに対して段落分割を行った．学習データは，Left-to-Right 型の HMM でモデル化しやすいデータセットとランダムに話題が転換するランダム型のデータセットを準備した．前者を Left-to-Right モデルと呼び，後者をランダムモデルと呼ぶ．ニュース記事の素材の仕様は以下の通りである．

(1) ウェブニュースの記事のジャンル：6 ジャンル (社会，経済，科学，娯楽，スポーツ，国際)
(2) 各ジャンル平均 263 記事づつ，計 1576 記事
(3) 1 つの記事の平均長 (単語数)：491 単語
(4) 記事の最小長および最大長：最小 390 単語，最大 5859 単語
(5) 記事集合全体で使われている単語の種類：9753

この記事データを以下の 2 通りの方法で 4 記事づつ結合し評価テキストを作成した．

(1) Left-to-Right モデルの評価テキスト
まず，6 つのジャンルに順番を定め固定する．1 つの評価テキストはこの 6 ジャンルのうち 2 ジャンルをランダムに選び削除した 4 つのジャンルからそれぞれランダムに選んだ 4 つの記事を，先に固定した順番で並べる．

(2) ランダムモデルの評価テキスト
すべての記事からランダムに 4 つを選び，結合する．

評価テキストは，上記 2 つの方法それぞれについて，100 テキストを 5 セット作成し，各セットごとに 40 テキストを学習データ，残り 60 テキストを評価データとした．以後，Left-to-Right モデルにより作成した評価テキストのセットを，ltr1, ltr2, ..., ltr5 と呼び，ランダムモデルにより作成した評価テキストのセットを data1, data2, ..., data5 と呼ぶ．

さらに，日を変えて収集したニュース記事から，以下のようなランダムモデルのデータを作成した．

(1) ウェブニュースの記事のジャンル：5 ジャンル (社会，国内，国際，娯楽，スポーツ)
(2) 各ジャンル平均 1493 記事づつ，計 7467 記事
(3) 1 つの記事の平均長 (単語数)：301 単語
(4) 記事の最小長および最大長：最小 14 単語，最大 2501 単語
(5) 記事集合全体で使われている単語の種類：21419

この記事データから，ランダムに 10 記事を選び結合したものを 1 つの評価テキストとする．学習データとして，評価テキストを 100 テキスト準備し，さらに別の 100 テキストを評価データとしたもを 1 セットとする．4 セットのランダムモデルのデータを作成し，それぞれ yar1, yar2, yar3, yar4 と呼ぶ．

3.2 評価方法と結果

評価テキスト ltr1, ltr2, ..., ltr5, data1, data2, ... , data5 および yar1, yar2, ..., yar4 に対してそれぞれの学習データを用いて HMM を構成する．その後，得られた HMM を用いて，評価データを段落分割し，分割位置の正しさを評価する．評価は，以下の値を比較した．

- 分割位置の完全一致に関する精度と再現率および F 値
- 前後 1 文のずれを許容した分割位置一致の精度と再現率および F 値
- 正しいジャンルに分類されている文章の割合 (分類率)

比較対象として，1 単語を出力記号とする従来の HMM による分割性能を示す．従来法は，最適状態遷移系列を求めたときに 1 文の中で最も多く留まった状態を文全体の状態と判定した．

表 1 に Left-to-Right モデルによる評価データ (ltr1,..,ltr5) の結果を示す．Left-to-Right モデルは従来型の HMM により効率良く分割できるテキストである．従来手法では，F 値でおよそ 0.4 の性能が得られていることがわかる．精度と再現率のバランスも，従来手法においてどちらもおよそ 0.4 である．分割位置の前後を分割候補として推定した場合も正解とした場合，F 値でおよそ 0.5 の向上が従来手法ではみられる．一方，提案手法では従来手法にくらべ，完全一致の場合でおよそ 0.05，前後を許容するとおよそ 0.1 の F 値の向上がみられた．これは，本手法を適用することにより，推定全体のばらつきが，より正解の分割位置に近い位置に偏ったと言える．提案手法における精度と再現率のバランスも従来手法と比べて劣るものではなく，安定してる．さらに従来手法においては，ltr5 に関する推定が他のデータに比べて良く，突出した値となっているが，本手法での性能では突出しておらず安定している．評価データの文それぞれが正しい段落として認識されたかが見える分類率においても提案手法の優位性がみられる．

次に，ランダムモデルによる評価データ (data1,..,data5) の結果を表 2 に示す．この評価データでは，従来手法の弱点であるランダムな話題の移り変わりに対して，本手法の性能向上が顕著である．従来手法では，再現率が極端に高く，精度が低い値となっている．これは，多くの分割位置を推定した場合に起こる現象である．これに対して本手法では，ランダムに話題が移り変わる場合でも精度と再現率のバランスが崩れにくく，およそ 0.4 の精度とおよそ 0.6 の再現率を保っている．正解の計測方法，データセットの違いのすべてについて，精度および F 値は，提案手法の性能が従来手法を上回っている．分類率に関しては，Left-to-Right モデルの場合と大差なく，従来手法が 0.6 台前半なのに対して，提案手法では，0.73 の分類正答率となっている．このことからも従来手法では，より細かな段落分割が行われ，性能低下を招いていると言え，提案手法が頑強であることが示された．

表 3 に別ソースのランダムモデルによる評価データ (yar1,..,yar4) の結果を示す．この場合も，従来手法では再現率が高く，精度が非常に低い傾向となり，多くの分割推定が行われている．対して提案手法では，精度と再現率のバランスの崩れが従来手法よりも小さい．さらに，分類率についても data1,..,data5 におけるランダムモデルの評価データと同じく，従来手法と提案手法において大きな差はなく，従来手法の細かい分割が見て取れる．

以上，Left-to-Right モデル，ランダムモデルどちらについても提案手法の優位が認められ，F 値の向上のみでなく，推定された段落が提案手法の方がより大きな段落となっていることがわかった．これにより提案手法は，従来手法の欠点であるランダムモデルに対する性能向上が望める手法であると言える．

Vol.2012-MPS-87 No.16
2012/3/2

4. おわりに

テキストセグメンテーションを HMM を用いて行う手法について，文の識別を行う識別器を状態に持つ HMM を提案し，評価実験において従来手法よりも高性能であることを示した．提案手法における HMM では，状態が段落を表す点は従来手法と同じだが，テキストを識別する過程で，単語ユニグラムによる段落のモデル化ではなく，その段落にふさわしい文章であるかを判定する識別器により段落をモデル化する．識別器として，ナイーブベイズによる二値分類器を考えると，HMM 全体のパラメータ学習に Baum-Welch アルゴリズムが利用でき，学習データから提案手法の HMM が構成できることを示した．その結果，Left-to-Right モデルによるテキストデータおよび，ランダムモデルによるテキストデータいずれの場合においても，従来手法より高性能であることが評価実験から確かめられた．特に，ランダムモデルにおける性能向上は，分割推定位置がより少なく，正解に近い位置に分布していることが確かめられた．今後の課題として，識別器をナイーブベイズ以外のものとした場合が挙げられる．

参 考 文 献

1) Hearst, M. A.: Texttiling: segmentaing text into multi-paragraph subtopic passages, Computaional Linguistics, Vol. 23, pp.33-64 (1997)
2) McCallum, A., Freitag, D., Pereira, F.: Maximum entropy markov models for information extraction and segmentation, Proc. of ICML'00, pp.591–598 (2000)
3) Ostendorf, M., Digalakis, V. V. and Kimball, O. A.: From HMM's to segment models: a unified view of stochastic modeling for speech recognition, IEEE Transactions on speech and audio processing, Vol. 4, No.5, pp.360–378 (1996)
4) Yamron, J.P., Carp, I., Gillick, L., Lowe, S., van Mulbregt, P.: A hidden markov model approach to text segmentation and event tracking, Proc. of IEEE conf. on Acoustics, Speech and Signal Processing, vol.1, pp.333-336 (1998)
5) 但馬康宏，北出大蔵，中林智，藤本浩司，小谷善行：HMM とテキスト分類器による対話の段落分割，情報処理学会論文誌　数理モデル化と応用，vol.2, no.2, pp.70–79 (2009)
6) 長野雄，鈴木基之，牧野正三：HMM を用いた複数 n-gram モデルによる言語モデルの構築，情報処理学会研究報告 SLP 40-26, pp.151–156 (2002)

Vol.2012-MPS-87 No.16
2012/3/2

表 1 Left-to-Right モデルに対する精度，再現率，F 値 および 分類率

		精度						再現率						F 値					
完全一致		ltr1	ltr2	ltr3	ltr4	ltr5	平均	ltr1	ltr2	ltr3	ltr4	ltr5	平均	ltr1	ltr2	ltr3	ltr4	ltr5	平均
	従来法	0.425	0.400	0.450	0.365	0.579	0.444	0.310	0.329	0.351	0.328	0.629	0.389	0.358	0.361	0.394	0.345	0.603	0.412
	提案手法	0.464	0.418	0.478	0.528	0.481	0.474	0.422	0.393	0.421	0.441	0.436	0.423	0.442	0.405	0.447	0.481	0.457	0.446
前後許容		ltr1	ltr2	ltr3	ltr4	ltr5	平均	ltr1	ltr2	ltr3	ltr4	ltr5	平均	ltr1	ltr2	ltr3	ltr4	ltr5	平均
	従来法	0.450	0.418	0.500	0.425	0.637	0.486	0.331	0.346	0.393	0.386	0.692	0.430	0.381	0.378	0.440	0.405	0.663	0.453
	提案手法	0.571	0.505	0.593	0.641	0.587	0.579	0.522	0.482	0.525	0.538	0.536	0.521	0.546	0.493	0.557	0.585	0.561	0.548

分類率						
	ltr1	ltr2	ltr3	ltr4	ltr5	平均
従来法	0.567	0.545	0.600	0.647	0.828	0.637
提案手法	0.768	0.761	0.751	0.746	0.772	0.760

表 2 ランダムモデルに対する精度，再現率，F 値 および 分類率

		精度						再現率						F 値					
完全一致		data1	data2	data3	data4	data5	平均	data1	data2	data3	data4	data5	平均	data1	data2	data3	data4	data5	平均
	従来法	0.127	0.106	0.085	0.091	0.098	0.101	0.728	0.769	0.742	0.700	0.732	0.734	0.216	0.186	0.153	0.161	0.174	0.178
	提案手法	0.323	0.322	0.321	0.312	0.337	0.323	0.575	0.481	0.418	0.483	0.408	0.473	0.414	0.386	0.363	0.379	0.369	0.382
前後許容		data1	data2	data3	data4	data5	平均	data1	data2	data3	data4	data5	平均	data1	data2	data3	data4	data5	平均
	従来法	0.167	0.131	0.113	0.119	0.120	0.130	0.957	0.951	0.961	0.943	0.933	0.949	0.285	0.230	0.202	0.211	0.213	0.228
	提案手法	0.391	0.407	0.385	0.375	0.411	0.394	0.681	0.617	0.513	0.574	0.515	0.580	0.497	0.490	0.440	0.453	0.457	0.467

分類率						
	data1	data2	data3	data4	data5	平均
従来法	0.673	0.637	0.598	0.615	0.610	0.627
提案手法	0.741	0.736	0.730	0.727	0.717	0.730

表 3 別のニュース記事集合から作成したランダムモデル評価データに対する性能

		精度					再現率					F 値				
完全一致		yar1	yar2	yar3	yar4	平均	yar1	yar2	yar3	yar4	平均	yar1	yar2	yar3	yar4	平均
	従来法	0.195	0.202	0.186	0.182	0.191	0.754	0.749	0.752	0.766	0.755	0.310	0.318	0.280	0.294	0.301
	提案手法	0.248	0.271	0.264	0.293	0.269	0.489	0.514	0.499	0.492	0.499	0.329	0.355	0.346	0.370	0.350
前後許容		yar1	yar2	yar3	yar4	平均	yar1	yar2	yar3	yar4	平均	yar1	yar2	yar3	yar4	平均
	従来法	0.242	0.253	0.236	0.226	0.239	0.938	0.953	0.960	0.948	0.950	0.385	0.400	0.379	0.365	0.382
	提案手法	0.360	0.372	0.372	0.419	0.381	0.718	0.707	0.722	0.694	0.710	0.479	0.487	0.491	0.522	0.495

分類率					
	yar1	yar2	yar3	yar4	平均
従来法	0.708	0.750	0.707	0.721	0.722
提案手法	0.715	0.738	0.715	0.752	0.730