ヒューマンインタフェース 53－8
（１９９４．３．１０）

複合化候補列類似度による分類

高橋邦夫，天沼博，熊谷憲二，鈴木光重，金子真輝，栗原雅明

神奈川大学工学部電気工学科

〒２２１横浜市神奈川区六角橋３－２７－１

分類は認識の前処理として有効である。また完全な認識結果が得られるとは限らないのでそのような場合分類は不可避である。この場合，平均個数が少なく，１個率の高い分類法が望ましい。

分類法としては，候補列の正解を含む順位までを候補として採用する類似度差による分類法を提案する。更に，分類に用いる類似度としては，候補列を求めた類似度と候補列に基づいた類似度との複合化候補列類似度を用いることを提案する。

この複合化候補列類似度を用いる類似度差分類法によって候補列を求めた類似度の複合化しない類似度の順位を用いる方法に比べ平均個数が減少することを示す。

Classification by Complex Candidates Series Similarity - Solution Preservation

Kunio Takahashi　Hiroshi Amanuma　Kenji Kumagai
Mitsushige Suzuki　Masaki Kaneko　Masaaki Kurihara

Department of Electrical Engineering
Kanagawa University

3-27-1 Rokkakubashi Kanagawa-ku Yokohama 221 JAPAN

In recognition of hand-written Chinese character, pre-classification is effective. It has been proposed to the method that determines candidate from the top to the value N of similarity. The purpose of the method is increasing kinds of character that has only one candidate and decreasing the average number of candidates. Now, we present the classification method that determineds candidate from the top to the predetermined value of complex candidate series similarity which is determined by learning. By this similarities and classification method, we can decrease the average number of candidates.

1．まえがき

分類[1]は認識の前処理として有効である。また完全な認識結果が得られるとは限らないのでそのような場合分類は不可避である。この場合，平均個数が少なく，1個率の高い分類法が望ましい。

分類法としては，候補列の正解を含む順位までを候補として採用する類似度差による分類法を提案する。更に，分類に用いる類似度としては，候補列を求めた類似度と従来用いられることのなかった候補列に基づいた類似度との複合化候補列類似度を用いることを提案する。

この複合化候補列類似度を用いる分類法によって候補列を求めた類似度の複合化しない類似度のグループ化分類結果の順位を用いる方法に比べ平均個数が減少することを示す。

従来，分類において，候補列を求めた類似度の上位を採用する手法が採用されている。しかし，本文では，必ず正解を含む条件を設定し，その条件を満足するまで候補として採用する。このような正解維持分類法により，平均個数を減少させることができる。また，候補列自身情報を有し，候補列によって求められる類似度によって類似度を複合化し，更に分類性能を増大させる。従来，候補列に基づく情報の利用はされなかった。

すなわち，本文では，候補列情報に基づく複合化類似度を提案し，更に，候補列順位1位との類似度差による分類手法をも提案し用いる。これによって，分類性能が候補列を求めた類似度の高順位を用いる手法に比べ向上することを示す。

2．諸定義

類似度をAとし，その類似度に基づく，候補列が求められているものとする。この候補列の候補列情報による候補列類似度をKとする。AとKの複合化類似度をAKまたは，AKCとする。

類似度Cは真の値からの距離を示すものとする。従って，小さな値が良いものとなる。類似度Sは大であれば良い値となるものである。

字種i，サンプル番号ℓの学習候補列を$K_\ell(i,j)$とする。字種jが存在すれば$K_\ell(i,j)=1$，存在しなければ$K_\ell(i,j)=0$とする。一般の候補列を$K(i)$とする。字種iが存在すれば$K(i)=1$とし，存在しない場合は$K(i)=0$とする。

類似度Aによる候補列が求められている場合，$C_{ij\ell}$は字種j，サンプル番号ℓの入力に対する字種iの類似度A，あるいはK，AKである，同様に，類似度Aによる候補列が求められている場合，C_iは字種iの類似度A，あるいはK，AKである。

$S_{ij\ell}$，S_iも同様に定義される。

分類率とは採用した候補列に正解を含む率である。1個率とは候補が1個になった率である。

3．候補列類似度

3.1候補列辞書

辞書$C^{(L)}(i,j)$は字種iの学習候補列$K_\ell(i,j)$のL位までに候補字種jが存在すれば$C_\ell(i,j)=1$，存在しない場合$C_\ell(i,j)=0$とする。

辞書$D^{(L)}(i,j)$は次のように求められる。

$$D^{(L)}(i,j)=\frac{w'_{ij}-\overline{w_{ij}}}{\delta w_{ij}}\times 10+50 \tag{1}$$

$$w'_{ij}=\sum_{\ell}K_\ell(i,j) \tag{2}$$

$$\overline{w_{ij}}=\sum_{\ell}w'_{ij\ell}/J \tag{3}$$

$$\delta w'_{ij}=\sqrt{\sum\left(w'_{ij}-\overline{w_{ij}}\right)^2/J} \tag{4}$$

また，Jは候補列の候補個数である。また，$\overline{S}$はSの平均値を示す。

字種iの辞書$E^{(L)}(i,j)$は次のようにして$D^{(L)}(i,j)$と$C^{(L)}(i,j)$から構成する。すなわち，$D^{(L)}(i,j)$の上位10位までのjに対しては$3C^{(L)}(i,j)$として11位より40位までに対しては$2C^{(L)}(i,j)$とし，それ以下に対しては$C^{(L)}(i,j)$とする。

3.2候補列類似度

候補列情報に基づく候補列類似度$K_i^{(L)}$を次のよ

うに定める。

$$K_i^{(L)} = \sum_j E^{(L)}(i,j) \times K(j) + w(t) \tag{5}$$

ここで，tは順位を示す。また

更に，次のように偏差値化する。

$$KN_i^{(L)} = \frac{K_i^{(L)} - \overline{K_i^{(L)}}}{\sqrt{\sum_j \left(K_i^{(L)} - \overline{K_i^{(L)}}\right)^2 / J}} \times 10 + 50 \tag{6}$$

$$\overline{K_i^{(L)}} = \sum_i K_i^{(L)} / J \tag{7}$$

更に，候補列類似度 $KL_i^{(L)}$ を次のように構成する

$$KL_i^{(L)} = \frac{\overline{K_i^{(L)}} - K_i^{(L)}}{\sqrt{\sum \left(K_i^{(L)} - \overline{K_i^{(L)}}\right)^2 / J}} \times 10 + 50 \tag{8}$$

$$= -KN_i^{(L)} + 100 \tag{9}$$

4．複合化候補列類似度

候補列類似度Kと候補列を求めた類似度Aとの複合化候補列類似度AKを次のように定める。

$$AK_i^{(L)} = \gamma K_i^{(L)} - \alpha \frac{A_i}{10} \sqrt{\sum \left(K_i - \overline{K_i}\right)^2 / J} + \sum_i K_i / J \tag{10}$$

また，$AK_i^{(L)}$は次のようにも構成できる。

$$AK_i^{(L)} = \mu KN_i + \beta \frac{A_i - \overline{A_i}}{\sqrt{\sum \left(A_i - \overline{A_i}\right)^2 / J}} \times 10 + 50 \tag{11}$$

$$\overline{A_i} = \sum_i A_i / J \tag{12}$$

更に次のように，候補列を求めた類似度Aと候補列類似度Kとの複合化候補列類似度 $AKC_i^{(L)}$ を偏差値を用いて構成する。

$$AKC_i^{(L)} = \mu A_i + \beta KL_i^{(L)} \tag{13}$$

5．類似度差分類法[6]

必ず正解を含む条件を設定する。その条件を満足する字種まで候補として採用する。

5.1　1位字種別パラメータ分類法

次の条件を満足する場合，字種jは候補として採用する。

$$S_j \geq S_i - \delta T_i \tag{14}$$

$\delta \geq 1.0$

ただし，T_iは次のように求められる。

$$T_i = \max_{j,\ell} \left(S_{ij\ell} - S_{jj\ell}\right) \tag{15}$$

ただし，字種iは候補列順位1位とする。

更に，本分類法を次のように構成できる。次の条件を満足する場合，字種jは候補として採用する。

$C_i \leq C_j + \delta T_i$, $\delta > 1.0$

ただし，T_iは次のように求められる。

$$T_i = \max_{j,\ell} \left(S_{ij\ell} - C_{jj\ell}\right)$$

5.2　類似度差統一パラメータ分類法

次の条件を満足する場合，字種jは候補として採用する。

$$S_j \geq S_i - \delta T \tag{16}$$

ただし，Tは次のように求められる。

$$T = \max_{i,j,\ell}\left(S_{ij\ell} - S_{jj\ell}\right) \quad (17)$$

ただし，字種iは候補列順位１位とする。

また，簡単のためとδTをパラメータPとして分類する。これを類似度差統一パラメータ分類法とする。

$$S_i \geq S_i - P \quad (18)$$

更に，類似度差統一パラメータ分類法を次のように構成できる。次の条件を満足する候補jは採用する。すなわち，次の条件を満足する場合，字種jは候補として採用する。

$$C_j \leq C_i + P \quad (19)$$

6．計算例

6.1 利用したデータ

ＥＴＬ－８を用いる。奇数サンプルを学習データとした。

6.2 候補列データ

類似度$D^{(3)}$を用いて，30位までを計算した結果を表１に示す。これをデータ１とする。それを用いて類似度$DS_2Z^{(3)}$により候補列を求めた結果を表2に示す。これをデータ２とする。

更に，類似度$D_2S_2ZM^{(7)}$により候補列を求めた結果を表3に示す。これをデータ３とする。

6.3 類似度差統一パラメータによる分類

Pをパラメータとし，式(5)において，$w(1)=12$, $w(2)=6$, $w(3)=5$, $w(4)=4$, $w(5)=3$, $w(6)=1$とする。

表1　類似度Dによる順位率（データ１）

順　位	1	5	10	20	30
分類率[%]	90.536	97.843	98.604	99.247	99.601

表2　類似度DS_2Zによる順位率（データ２）

順　位	1	5	10	20	30
分類率[%]	96.331	99.348	99.564	99.633	99.651

表3　類似度D_2S_2ZMによる順位率（データ３）

順　位	1	5	10	20	30
分類率[%]	96.911	99.434	99.596	99.638	99.648

る。

まず，候補列データとしてデータ２を用いて分類を行う。この場合，分類に用いる複合化類似度$AK_i^{(L)}$としては，候補列類似度$K_i^{(L)}$と類似度DS_2ZをA_iとして式(10)により求める。分類結果を表4，表5，表6に示す。

次に候補列データとしてデータ１を用いる分類を行う。この場合，分類に用いる複合化類似度$AK_i^{(L)}$として，候補列類似度$K_i^{(L)}$と類似度DをA_iとして式(10)により求める。表7に結果を示す。

類似度D_2S_2ZMをA_iと，その候補列類似度Kを(10)により複合化し複合化候補列類似度D_2S_2ZMKを構成し分類した結果を表8に示す。

表4　複合化候補列類似度DS_2ZKによる分類(α=1.0,γ=1.0)

P	平均個数	分類率[%]	1個率[%]
6	1.187	99.000	87.641
8	1.427	99.342	75.476
10	1.860	99.510	58.127
12	2.974	99.592	38.272
14	3.652	99.637	20.531

表5　複合化候補列類似度DS_2ZKによる分類(α=0.0,γ=1.0)

P	平均個数	分類率[%]	1個率[%]
12	2.942	99.382	23.235
14	4.054	99.498	10.547
16	5.507	99.570	3.798
18	7.309	99.603	1.050
20	9.478	99.620	0.240

表6　複合化候補列類似度DS_2ZKによる分類(α=1.0,γ=0.0)

P	平均個数	分類率[%]	1個率[%]
6	1.432	99.247	81.550
8	1.810	99.455	71.797
10	2.411	99.561	59.545
12	3.318	99.612	59.545

6.4 グループ化分類法[4]の結果

グループ化分類により1個率の増加が可能である。

まづ，類似度DS_2Zによる候補列データ2によりグループ化分類[4]を行う。この場合，複合化類似度としては，候補列類似度$KL_i^{(L)}$とDS_2Zを類似度Aとして(13)より複合化類似度DS_2ZKCを求める。グループ化分類結果を表9に示す。

グループ化分類法による結果の順位と分類率を表10に示す。

更に，データ2のグループ化分類法の結果を表11に示す。

更に，Dによるグループ化分類の結果を表12，13に示す。

データ3のグループ化分類結果を表14に示す。

データ2に対する複合化候補列類似度による統一パラメータ類似度差分類の結果（表4）とグループ化分類との共通候補とした結果を表15に示す。

データ3に対する複合化候補列類似度による類似度差分類の結果（α=γ=0.5）とグループ化分類

表7　複合化候補列類似度DKによる分類

α	γ	P	平均個数	分類率[%]	1個率[%]
1.0	1.0	18	6.095	99.493	7.245
1.0	1.0	20	8.337	99.579	2.353
0.0	1.0	20	6.874	99.391	15.790
0.0	1.0	18	5.245	99.230	23.818
1.0	0.0	16	6.719	99.420	3.234
1.0	0.0	18	9.238	99.523	0.857
1.0	0.0	20	12.259	99.586	0.178

表8　複合化候補列類似度D_2S_2ZMによる分類(α=0.8, γ=0.2)

P	平均個数	分類率[%]	1個率[%]
1	1.017	97.665	98.539
2	1.048	98.341	96.244
3	1.099	98.846	93.114
4	1.176	99.135	89.132
5	1.291	99.359	83.834
6	1.462	99.471	76.859
7	1.705	99.547	68.175
8	2.042	99.587	58.133

表9　複合化類似度DS_2ZKCによるグループ化分類

複合化パラメータ		分類率[%]	平均個数	1個率	δ	γ	n
μ	β						
0.8	0.2	99.516	2.617	60.607	0.8	10	4

表10　複合化類似度DS_2ZKCによるグループ化分類結果(μ=0.8,β=0.2)

順位	平均個数	分類率[%]
1	1.000	96.426
2	1.366	98.560
3	1.603	99.048
4	1.779	99.235
5	1.919	99.333
6	2.059	99.398
7	2.199	99.440
8	2.338	99.469
9	2.478	99.495
10	2.618	99.516

表11　類似度DS_2Zによるグループ化分類結果

順位	平均個数	分類率[%]
1	1.00	96.331
2	1.394	98.464
3	1.659	98.977
4	1.862	99.190
5	2.030	99.304
6	2.198	99.384
7	2.366	99.428
8	2.534	99.464
9	2.701	99.500
10	2.870	99.515

表12　Dによるグループ化分類結果

分類率[%]	平均個数	1個率[%]	δ	γ	n
99.601	21.982	19.052	0.8	10	4

結果表14との共通候補とした分類結果を表16に示す。

7．考察

データ3による類似度差分類結果（表8または表16）と候補列を求めた類似度のみによるグループ化分類（表14）の結果を比較すれば，類似度差分類の効果が明らかである。同一分類率において約35%の平均個数の減少が達成されている。またデータ1によるグループ化分類結果の順位により分類（表12，13）と複合化候補列類似度による分類の結果（表7）を比較すれば，候補列を求めた類似度のみによるグループ化分類結果の順位のみの分類に比べ複合化候補列類似度による類似度差分類法の効果が明らかである。約50%～70%の平均個数の減少が達成されている。

複合化候補列類似度の効果は表7におけるα=1.0，γ=1.0の結果とα=1.0，γ=0.0の結果とを比較すれば明らかである。また，複合化候補列類似度DKによるα=1.0，γ=1.0，P=20の結果（表7）は複合化候補列類似度D_2S_2ZMKによる分類のためのデータとして利用できる。

表8による結果は高い分類率99.359%，少ない平均個数1.291個，高い1個率83.834%を有し，認識の前処理データとして有効である。

8．あとがき

表13　Dによるグループ化分類結果

順位	平均個数	分類率[%]
1	1.000	90.576
2	1.814	95.207
3	14.805	99.409
4	15.523	99.436
5	16.240	99.457
6	17.676	99.508
7	18.393	99.530
8	19.111	99.550
9	21.264	99.590
10	21.982	99.301

類似度差による統一パラメータ分類法の効果が明らかにされた。

候補列を求めた類似度の順位を用いる分類法に比べ，複合化候補列類似度の類似度差による統一パラメータ分類法の効果が明らかにされた。

ＥＴＬ－８作成された電総研関係者に感謝いたします。

表14　類似度D_2S_2ZMによるグループ化分類結果

順位	平均個数	分類率[%]
1	1.000	96.844
2	1.376	98.534
3	1.630	99.015
4	1.820	99.217
5	1.969	99.344
6	2.084	99.406
7	2.167	99.446
8	2.248	99.480
9	2.327	99.495
10	2.406	99.506

表15　複合化候補列類似度DS_2ZKによる結果（表4）と複合化類似度によるグループ化分類結果（表9）との共通候補を候補とする

P	平均個数	分類率[%]	1個率[%]
6	1.166	98.950	89.423
7	1.240	99.143	85.961
8	1.330	99.275	82.533
9	1.431	99.353	79.442
10	1.544	99.415	76.824

表16　複合化候補列類似度D_2S_2ZMKによる結果とグループ化分類結果（表14）との共通候補を候補とする

P	平均個数	分類率[%]	1個率[%]
6	1.125	99.000	91.475
7	1.286	99.320	83.389
8	1.385	99.392	79.578

参照文献

(1) Mori S., Yamamoto K. and Yasuda M. : "Research on Machine Recognition of Handprinted characters", IEEE Trans. Pattern Anal. Mach. Intell., PAMI-6, 4 (July 1984).

(2) T.Saito, H.Yamada, K.Yamamoto, S.Mori: "An Analysis of Handprinted Character Data Base V-Evaluation of KYOUIKU-KANJI Characters by Pattern Matching Approach", Bullelin of ETL Vol.-45, No.-1 (1981)

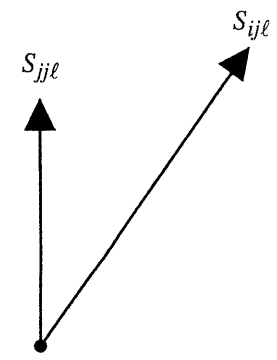
(3) 高橋邦夫，天沼　博，加藤弘之：" 手書き漢字の学習パラメータによる分類ー構造化パターンマッチング等によるー"，信学論(D-II)，J75-D-II, 5, pp.674-675(1992-03)

(4) 高橋邦夫，天沼　博，加藤弘之：" 学習グループ化による手書き漢字の分類"，信学論(D-II)，J75-D-II, 9, pp.1626-1627(1992-09).

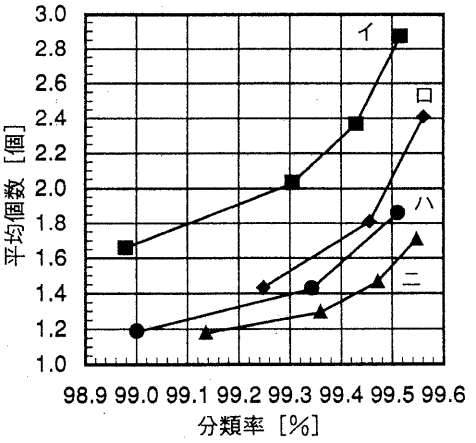
(5) 高橋邦夫，天沼　博，金子真輝，鈴木光重，栗原雅明，熊谷憲二，「学習分類法」，電子情報通信学会パターン認識・理解研究会，PRU93-101，1993-12

(6) 高橋邦夫，天沼　博，小林賢一，熊谷憲二：" 最高位候補列パラメータによる候補列マッチング分類"，情報処理学会第47回全国大会，2L-7，1993-10

(7) 高橋邦夫，天沼　博，山口　忍，寺尾英幸：" 形態情報によるグループ化・候補列マッチング分類"，1993信学春季全大，D-544



図．1　類似度差分類法の説明図



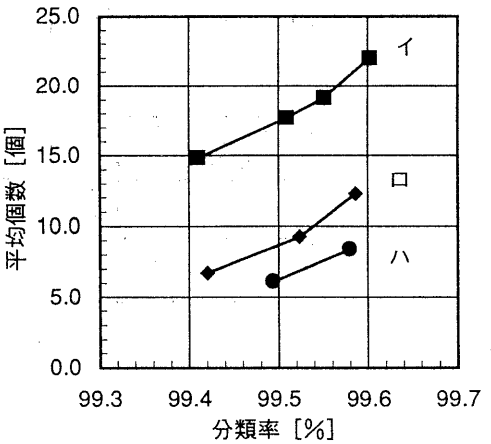
図．2　分類結果

イ．グループ化分類結果（表11）

ロ．類似度DS_2ZKによる類似度差分類結果（表6）

ハ．複合化候補列類似度DS_2ZKによる類似度差分類結果（表4）

ニ．複合化候補列類似度D_2S_2ZMKによる分類結果（表8）



図．3　分類結果

イ．グループ化候補列分類結果（表13）

ロ．類似度Dによる類似度差分類結果（表7）

ニ．複合化候補列類似度DKによる類似度差分類結果（表7）

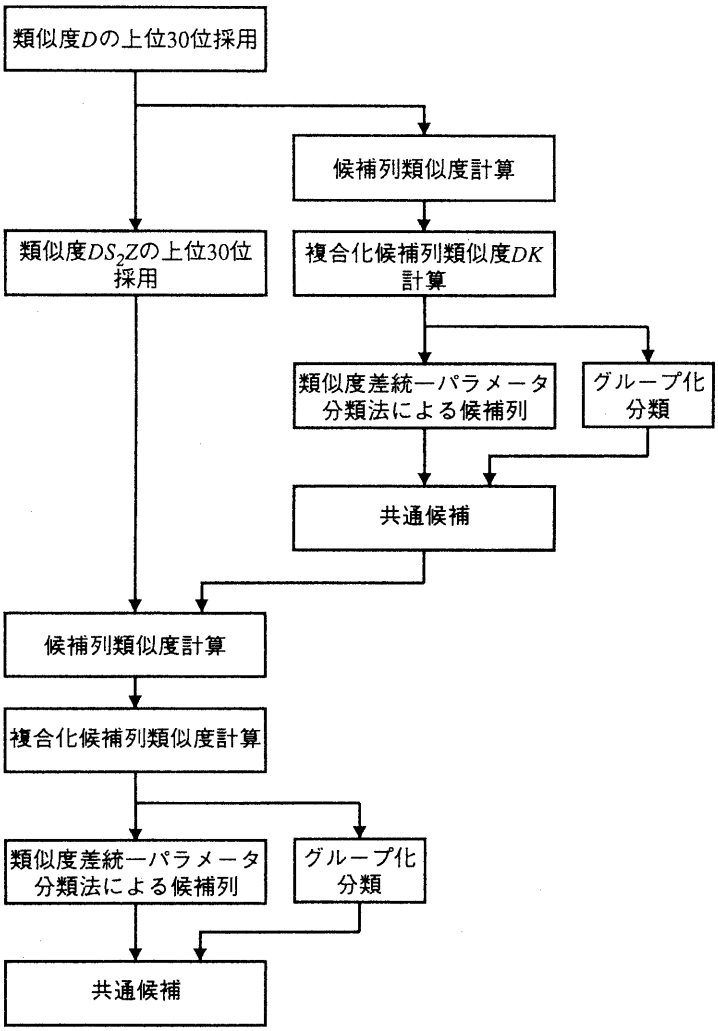


図. 4　本手法概要の説明図