

数量化理論第Ⅰ類を用いた手書き類似文字の詳細識別

Discrimination of Handprinted Similar Characters
by Quantification Theory Type-I

太田 勝[†] 西脇 大輔[†] 伊藤 彰義^{†*} 川西 健次^{†*}
Masaru OOTA Daisuke NISHIWAKI Akiyoshi ITOH Kenji KAWANISHI

日本大学 [†]大学院理工学研究科 ^{†*}理工学部電子工学科

NIHON Univ. , [†]Graduate School ^{†*}College of Sci.& Tech. ,Dept. of Electronic Engineering

Abstract In our previous report , we have proposed a method of discriminating of printed similar characters by quantification theory type-III . This method can reduce the discriminating time and the storage capacity of a standard patterns .

Handprinted characters have more fluctuation than printed characters . So this method is inapplicable to handprinted characters .

In this report , we propose a discriminating method of handprinted similar characters by quantification theory type-I .

あらまし 漢字文字の認識において効率良く有効な類似文字の識別が望まれる。以前我々は印刷類似文字の識別特徴を抽出(マスク作り)する方法として、数量化理論第Ⅲ類を用いることを提案した。本報告では手書き類似文字の識別を考える。しかし、手書き文字の識別特徴抽出法として数量化理論第Ⅲ類を適用した場合、適切な識別特徴抽出を行うことが難しい。これは手書き文字が印刷文字よりも大きな変動を含み、数量化理論第Ⅲ類がこの変動を際立たせるためである。我々は外的基準のある数量化法を用いることにより手書き類似文字の変動に対処した識別特徴抽出ができ、識別能力を向上できると考え、数量化理論第Ⅰ類を用いることを検討した。また数量化理論第Ⅰ類は識別操作に用いる特徴項目数に制限が加わるため、特徴項目選択の一手法として相関比による特徴項目選択の性質を検討した。この結果、数量化理論第Ⅰ類と相関比による特徴選択法を合わせ用いることにより、類似文字の識別特徴抽出が適切に行われ、効率がよく有効な類似文字の識別が行えることが分かった。

1. まえがき

最近、手書き漢字の認識にもマッチング法を用いるようになった。マッチング法は構造解析法に比べて認識のためのアルゴリズムが単純になるという利点がある。マッチング法を手書き漢字に適用するには、文字データから文字線の構造情報を反映した特徴ベクトルを抽出する方法を用いることが多い。

漢字文字の認識過程を考えると、大分類から詳細識別に至る階層的な分類過程を成すことが一般的である。このような分類過程を形成する理由は、認識対象文字種の増大に対処するためである。認識対象文字種の増大は類似文字の増大につながる。したがって詳細識別の過程では、類似文字の識別の問題は避けて通れない。またマッチング法を用いた場合、類似文字の識別能力向上に対する施策は難しいとされる¹⁾。これらのことから類似文字の識別を検討することは重要である。

さて、詳細識別の過程では、特徴抽出、識別という2つの操作を行うことになる。もちろん特徴抽出を大分類や中分類の過程と共用することも可能である。以前特徴抽出というと、文字データを直接識別空間に写像する操作をさしていた²⁾。しかし、文字データより構造情報を反映した特

徴ベクトルを抽出し、さらにこの特徴ベクトルを識別空間に写像する操作を行うことにより、2段階の特徴抽出操作を行うことになる。本報告ではこれらを区別するために、前者を構造特徴抽出、後者を識別特徴抽出と呼ぶことにする。

識別特徴抽出では、構造特徴抽出で抽出した文字特徴ベクトルの全ての要素を使う必要は、必ずしもない。特に今回のように識別の対象が類似文字に限定される場合は、この傾向が強くなる。この識別特徴抽出操作に使うベクトル要素を選び出す操作を特徴選択と呼ぶ。また識別特徴抽出に数量化理論第Ⅰ類を用いるためには特徴項目数を減らすことが必要になるため、今回の識別に、特徴選択操作は不可欠のものである。さらに、これらの識別特徴抽出や特徴選択は、文字認識システム構築時の辞書容量の削減や識別スピードの向上にも寄与する。

本報告では、特徴選択に相関比を用いた特徴選択法を、そして識別特徴抽出法に数量化理論第Ⅰ類を用いることにより、有効な類似文字識別が行えることを、いくつかの例により示す。

2. 類似文字集合

詳細識別における識別を検討する上で、この階層で識別する文字種（認識対象文字種）を設定する必要がある。本来、詳細識別を考える上で、この分類過程の上位の過程における分類結果を用いるのが順当な方法である。しかしこのような方法は、上位の分類の手法が変更になった場合に、適用できなくなるおそれがある。上位の分類方法が変更になっても適用できる詳細識別の手法を検討するために、識別が困難だと考えられる類似文字を詳細識別のモデルと位置づけ、類似文字の識別を試みた。

2.1 類似文字集合の作成

実験には電総研提供の手書き教育漢字データベース ETL-8B³⁾⁴⁾（160字／字種、64×63=4032点、各点1bit）を使用した。はじめに ETL-8B に付属する見本文字（ゴシック体）を63×63点からなる正方形枠いっぱいに入るように正規化し、次節で述べる外郭方向寄与度特徴⁵⁾を抽出した。この見本文字の特徴を用いて、文字相互のユークリッド2乗距離を求め、この距離が近くなった文字集合を類似文字集合とする。

さて今回は識別を2次元の平面散布図上で行うことを考える。平面散布図上の各象限にそれぞれのカテゴリを配置し、平面散布図の2方向の軸を識別境界線とすると、4つのカテゴリの識別が可能である。そこで4つの文字種からなる類似文字集合を作った。今回実験に用いた類似文字集合を表1に示す。

類似文字集合が決まったので、データベースより類似文字集合の手書き文字のデータを抜き出し、正方形枠いっぱいに入るように正規化し、外郭方向寄与度特徴を抽出した。この抽出した文字特徴データのうち、偶数番データセットを学習文字特徴データに、奇数番データセットを未知文字特徴データとして以下の実験に用いた。

表1 類似文字集合の例

類似文字集合	メンバー
No.1	回回回回
No.2	開開開開
No.3	自日白目
No.4	犬太太六

2.2 外郭方向寄与度特徴

ここで、構造特徴抽出で抽出する外郭方向寄与度特徴⁵⁾（Peripheral Direction Contributivity PDC特徴）について説明する。この特徴は萩田らによって提案されているパターンマッチングに基づく特徴で、文字線の複雑さ、文字線の方向、文字線の接続関係、文字線の相対位置関係の4種類の構造情報を反映したものである。

はじめに方向寄与度を算出する。方向寄与度は文字内の各黒点について4次元ベクトルで表される。図1に示す文字線の黒点の方向寄与度 d_o を

$$d_o = (d_{1o}, d_{2o}, d_{3o}, d_{4o}) \quad (1)$$

で表す。各要素 d_{mo} ($m=1,2,3,4$) は点から図1に示す8方向

に触手を伸ばして求まる黒点連結長 l_i ($i=1,2,\dots,8$) を用いて

$$d_{mo} = \frac{l_m + l_{m+4}}{\sqrt{4 \sum_{j=1}^4 (l_j + l_{j+4})^2}} \quad (2)$$

で定義される。

この方向寄与度を用いて外郭方向寄与度特徴は次のように算出される。文字を45°おきに8方向から走査し、横切る各文字線の最初に横切る輪郭点での方向寄与度を図1に示すように投影する。走査方向を t ($t=1,2,\dots,8$) とし、 n 本目 ($n=1,2,3$) に横切った文字線の方向寄与度 m 成分 ($m=1,2,3,4$) の任意の点 z における関数を $H_{tmn}(z)$ とおく。次に投影軸を7区間に等分割し、各区間内の $H_{tmn}(z)$ の値を平均して、外郭方向寄与度特徴が抽出される。この特徴の各要素を $P_{tmn}(k)$ ($k=1,2,\dots,7$) とおくと、外郭方向寄与度特徴ベクトル P は次のようなベクトルで表される。

$$P = (P_{111}(1), \dots, P_{843}(7)) \quad (3)$$

このようにして抽出した特徴ベクトルは672次元（文献5では1536次元）となる。

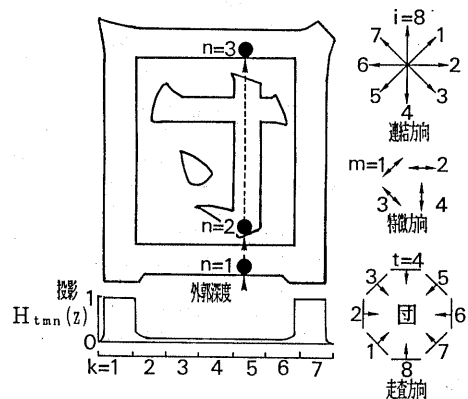


図1 外郭方向寄与度特徴

2.3 外郭方向寄与度の観察

外郭方向寄与度特徴は図1のように投影された量であるので、特徴ベクトルの値をみても直感的に特徴のイメージをつかむことが難しい。そこで、この特徴ベクトルを2次元文字データにあてはめ、特徴を表示観察してみる。図2に手書き文字データ「回」の外郭方向寄与度特徴を表示したものとする。この図における線の向きは外郭方向寄与度の方向を表し、線の密度は外郭方向寄与度の大きさを表している。ただし全ての要素を表示すると、図は非常に見づらいものとなるので、要素の大きいところだけを相対的に抽出表示し、外郭点でない画素や、対応する方向寄与度が小さい画素はドットとして表示してある。このように外郭方向寄与度を2次元文字パターンにあてはめ表示することで、特徴がどのように抽出されているか感覚的にとらえることができる。

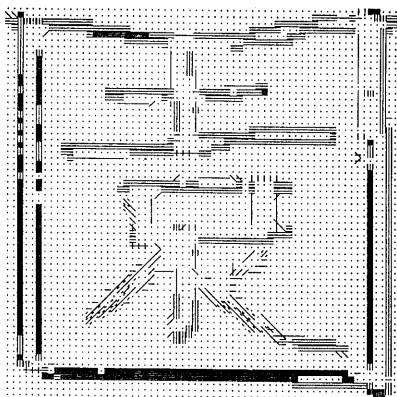


図2 外郭方向寄与度の観察
手書き文字“園”の例 PDC 672項目

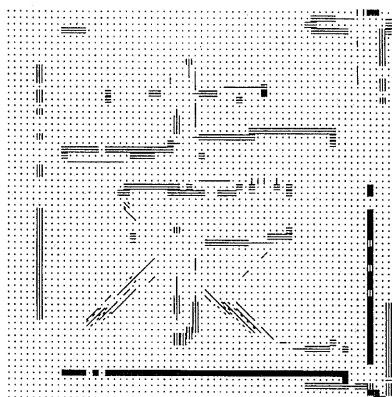


図3 特徴選択をした外郭方向寄与度の観察
手書き文字“園”の例 PDC 120項目

3. 相関比による特徴項目の選択

2次元文字データから抽出される構造特徴ベクトルには、識別に有効な特徴ベクトル要素（特徴項目）と識別に不適切な特徴要素が含まれていると考えられる。したがって、何らかの尺度を特徴ベクトル要素ごとに見出して、その尺度をもとに識別操作に使用するベクトル要素の選択（カテゴリ内ベクトル要素の変動が少ない要素を選択）を行えば、識別はより有効なものとなると考えられる。この章では、相関比による特徴項目の選択⁹⁾が類似文字の識別に及ぼす影響について検討する。

3.1 相関比

表2はn個のサンプル（各々はK個のカテゴリの何れかに属する）がR個の特徴項目に反応するデータの構造を示したものである。このデータにおいて、第j番目の特徴項目の全分散 $S_T(j)$ 及びカテゴリ間分散 $S_B(j)$ は

$$S_T(j) = \frac{1}{n} \sum_{i=1}^K \sum_{a=1}^n (\delta_{ia}(j) - \bar{\delta}_{..}(j))^2 \quad (4)$$

$$S_B(j) = \frac{1}{n} \sum_{i=1}^K n_i (\bar{\delta}_{i.}(j) - \bar{\delta}_{..}(j))^2$$

$$\text{ただし } \bar{\delta}_{..}(j) = \frac{1}{n} \sum_{i=1}^K \sum_{a=1}^n \delta_{ia}(j)$$

$$\bar{\delta}_{i.}(j) = \frac{1}{n_i} \sum_{a=1}^n \delta_{ia}(j)$$

となる。これにより第j番目の特徴項目の相関比 $\eta^2(j)$ は次の式で表される。

$$\eta^2(j) = \frac{S_B(j)}{S_T(j)} \quad (5)$$

式(5)より相関比が大きい特徴項目とはカテゴリ間分散が大きいことが分かる。また、全分散 = (カテゴリ間分散) + (カテゴリ内分散) という関係があるので、相関比が大きい特徴項目はカテゴリ内分散が小さい（カテゴリ内ベクトルの変動が少ないこと）ことも分かる。カテゴリの識別（パターンの認識）を行うために有効な特徴項目とは、カテゴリ間分散が大きく、カテゴリ内分散が小さい特徴項目

だと考えることが出来るので、相関比が大きい特徴項目ほどカテゴリ識別に有効な特徴項目であると考えることが出来る。したがって各特徴項目毎に相関比を算出し、この値が大きい特徴項目のみを選べば、カテゴリ識別の操作に有効な特徴項目の選択が行えると考えられる。

3.2 選択項目の観察

相関比による特徴項目選択によって、外郭方向寄与度特徴のどの部分が選択され、どの部分が削除されるかを検討する。はじめに類似文字集合No.1の学習文字特徴から各特徴項目ごとに相関比を算出する。図2に示した未知文字“園”の外郭方向寄与度のうち、学習文字から算出した相関比の大きな120項目の特徴項目に相当する特徴を選択する。選択した特徴を図3に示す。この類似文字集合は国構えから成り立っている。したがってこれらの文字の識別を考える上で、国構えの部分は必要ない。図3では識別に必要な国構え内部の特徴が多く選択されていることが分かる。このことから相関比により、文字の識別に必要な特徴項目の選択が行えることが分かった。

表2 分析操作に用いる外的基準のあるデータ構造の例

特徴項目	1	2	j.....	R	外的基準
カテゴリ	サンプル				
1	1				y_1
	2				
	⋮				
	α				
	⋮				
	n_1		$\delta_{1\alpha}(j)$		
2	⋮				y_2
⋮	⋮				⋮
i	⋮				y_i
⋮	⋮				⋮
K	⋮				y_K

3.3 距離による識別

この節では、数量化理論を用いた特徴抽出法の実験を行う前に、相関比による文字特徴の選択が文字の識別にどのような影響を及ぼすかを検討する。このため選択した文字特徴を用いて距離による文字の識別実験を行った。

はじめに使用する文字特徴の項目ごとに相関比を算出する。ここで相関比の大きな特徴項目だけを選択し識別の操作に使用する。選択した特徴項目から学習文字の平均ベクトルを求め標準特徴とする。これと未知文字の選択した特徴ベクトルとのユークリッド2乗距離を算出し、識別を行う。選択する項目の数をパラメータとして選択項目数と識別率の関係を考察する。

類似文字セットNo.1の識別結果を図4に示す。図の縦軸は識別率を、横軸は相関比により選択した特徴項目の数を示す。

図4では特徴項目数を減らしていても識別率が99.1%で変化しない部分が項目数96まで続く。ここでは識別の対象を類似文字としたため、文字の特徴は冗長なものとなっており、文字特徴の中には識別に不必要な（寄与しない）特徴項目が多数含まれていると考えられる。この部分では識別に寄与しない特徴項目の削除が行われているため、特徴項目数を変化させても識別率が変化しない部分が長く続くと考えられる。

識別率が変化しない部分を過ぎると、識別率は変動を繰り返しながら下降して行く。ここでは手書き文字を識別の対象としているため、各特徴項目には同一文字種であっても大きな変動がある。特徴項目が多い場合には、特徴項目をいくつも使用することでこの変動に対処し、高い識別率が得られたが、項目数を極端に少なくすると変動に対処できなくなるため識別率が低下すると考えられる。

図4で変動を繰り返しながら下降していくのは、特徴選択が必ずしも最適に行われていないことを示す。もし最適な特徴項目の選択が行えているならば、このような変動は起こらず、識別率は下降して行くだけである。しかし最適な特徴項目の選択を行うためには、1つの特徴項目だけに注目するのではなく、選択する特徴項目の組合せを考慮にいれなければならない。しかし組合せの数は膨大であり、全ての組合せについて識別率を求めるのは不可能である。また用いる学習文字の特徴と未知文字の特徴の傾向によっても選択の善し悪しは変わってくる。

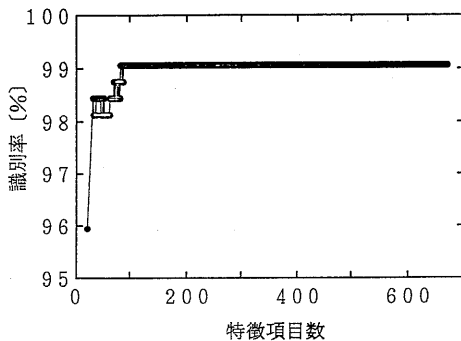


図4 ユークリッド2乗距離による識別
“回” “団” “国” “國” の例

図4の結果では全特徴項目を用いたときの識別率が96項目でも続けて得られているので、良好な特徴選択が行えたと考えられる。

3.4 3章まとめ

相関比による特徴項目の選択を行うことにより、類似文字の識別を考える上で、不必要な特徴項目を削除し必要な特徴項目を選択することができた。

4. 数量化理論第Ⅰ類による識別

以前我々は印刷類似文字の識別特徴抽出法として数量化理論第Ⅲ類を用いることを提案した⁸⁾⁹⁾。しかし手書き文字の識別特徴抽出法として数量化理論第Ⅲ類を適用した場合、適切な特徴抽出を行うことが難しい⁶⁾。そこで我々は外的基準のある数量化法を用いることにより手書き文字の変動に対処した識別特徴抽出ができ、識別の能力を向上できると考えた。

4.1 数量化理論

通常の変量解析法は比率尺度や間隔尺度で得られたデータを処理分析の対象とする。しかし実質科学のいろいろな分野では名義尺度や順序尺度で得られる質的なデータもある。ここで表3にデータの分類についてまとめておく。

それぞれの尺度に適用できると言われている統計量の関係を示す図5をみると、間隔尺度の統計量は比率尺度の統計量を包含していることが分かる。同様に順序尺度は間隔尺度を、名義尺度は順序尺度を包含している。これらの関係から、名義尺度に適用できる統計量から展開される解析法は、順序尺度、間隔尺度や比率尺度にも拡張が可能であると考えられる。

名義尺度や順序尺度に適用できる解析法として考案された手法が林の数量化法⁷⁾¹⁰⁾である。名義尺度や順序尺度から展開される数量化法は、名義尺度や順序尺度だけでなく、間隔尺度や比率尺度にも適用が可能であると考えられる。

表3 データの分類

データ	質的データ (定性的)	名義尺度
		順序尺度
	量的データ (定量的)	間隔尺度
		比率尺度

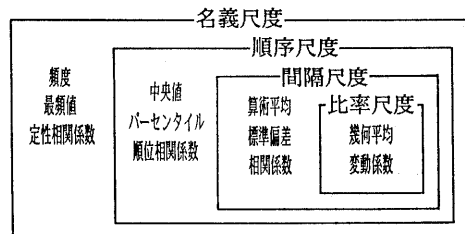


図5 統計量と尺度の関係

数量化法は表4に示すように、外的基準のある場合と外的基準のない場合とに分けられる。前者は数量化理論第Ⅰ類、第Ⅱ類、後者は数量化理論第Ⅲ類、第Ⅳ類と呼ばれる。このような考え方は数量化理論全体を見通すのに大変都合がよい。従来型の変量解析法と対応を考えると表5のようになる。数量化理論第Ⅰ類や第Ⅱ類は、名義尺度や順序尺度に適用できることから、重回帰分析や重判別分析を包含拡張する手法としてとらえることができる。数量化理論第Ⅲ類は従来型の変量解析法としては主成分分析法に類似しているが、完全に同一の手法であるとは言えない。しかしこれも、主成分分析法の分析尺度を拡張する手法として捕らえることができる。数量化理論第Ⅳ類は従来型の変量解析法にはない解析法である。

それぞれの数量化法は先に述べたように従来の代表的な3種類の変量解析法を包含すると共に、従来の変量解析法の見通しを良くする方法としてとらえることもできる。

4.2 数量化理論第Ⅰ類

数量化理論第Ⅰ類は各サンプルに対する多次元反応データ（特徴項目）から、外的基準の値を説明あるいは予測するための方法である。

いま表2のように1つの外的基準 y と R 個の特徴項目について、 n 個のサンプルが与えられたとする。 y_i の値を線形関数で予測するとき、 $\delta_{i.}(1), \delta_{i.}(2), \dots, \delta_{i.}(R)$ だけでは説明しきれない予測誤差の平方和 ε は次式で表される。

$$\varepsilon = \sum_{i=1}^n \sum_{j=1}^R \{ y_i - \sum_{j=1}^R a_j (\delta_{i.}(j) - \bar{\delta}_{..}(j)) \}^2 \quad (6)$$

$$\text{ただし } \bar{\delta}_{..}(j) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^R \delta_{i.}(j)$$

予測誤差の平方和を最小にする係数 $a_1, \dots, a_R$ は次の式によって算出できる。

$$a = V^{-1} S, \quad a = (a_j | j=1, \dots, R)^T \quad (7)$$

ただし

$$V = [V_{j,j'}]_{(R,R)}, \quad S = (S_j | j=1, \dots, R)^T$$

$$V_{j,j'} = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^R (\delta_{i.}(j) - \bar{\delta}_{..}(j)) \times (\delta_{i.}(j') - \bar{\delta}_{..}(j'))$$

$$S_j = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^R (y_i - \bar{y}_{..}) (\delta_{i.}(j) - \bar{\delta}_{..}(j))$$

$$\bar{y}_{..} = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^R y_i$$

予測（説明）するサンプルの反応データを $\delta'_{i.}(j)$ ($j=1, \dots, R$)とすると予測値（スコア） $\hat{y}$ は次のようになる。

$$\hat{y} = \sum_{j=1}^R a_j (\delta'_{i.}(j) - \bar{\delta}_{..}(j)) \quad (8)$$

この式を構造特徴空間から識別特徴空間の1つの成分軸への写像として用いる。2組の外的基準に対してそれぞれ式(7)(8)を適用しスコアを求めることで、識別平面への写像が行える。

表4 数量化理論の分類

		数量化理論
外的基準のある場合	外的基準が量的	第Ⅰ類
	外的基準が質的	第Ⅱ類
外的基準のない場合	サンプルの項目に対する反応から分析	第Ⅲ類
	サンプル同士の類似性から分析	第Ⅳ類

表5 数量化理論と従来型の変量解析法の対比

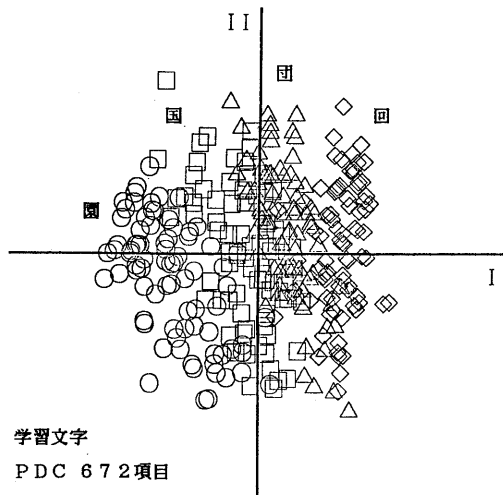
数量化理論	従来型の変量解析法
第Ⅰ類	= 重回帰分析法
第Ⅱ類	= 重判別分析法
第Ⅲ類	≈ 主成分分析法
第Ⅳ類	—

4.3 識別実験

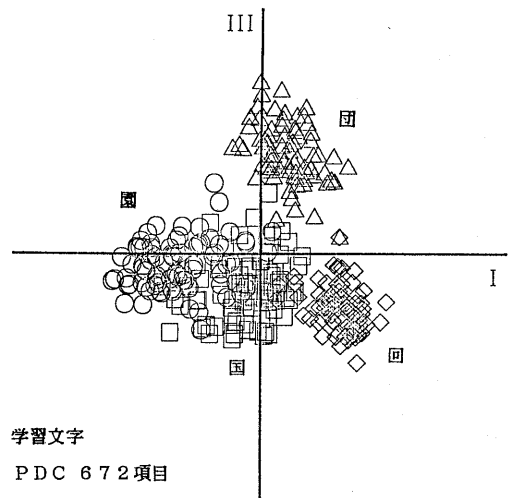
はじめに外的基準を決定するため、外的基準のない数量化法である数量化理論第Ⅲ類を用いて、学習文字の特徴ベクトルを分析する。数量化理論第Ⅲ類の数量の算出方法については、文献8、9を参照してもらいたい。

この分析結果を図6に示す。第Ⅰ、第Ⅱ主成分をプロットした図6(a)を見ると、第Ⅱ軸の各プロットは文字種に全く関係なくプロットされ、この軸は全く文字の識別には寄与しないことが分かる。第Ⅰ、第Ⅲ主成分をプロットした図6(b)の結果を見ると“団”と“国”は相反した位置に、“園”と“回”は相反した位置にプロットされていることが分かる。この位置関係を数量化理論第Ⅰ類の外的基準に用いる。

“国”、“園”の字種の外的基準を+1、“回”、“団”の外的基準を-1として式(7)を用いて、数量化理論第Ⅰ類の係数ベクトル a を求める。この値から式(8)を用いて学習文字に対する予測値と未知文字に対する予測値



(a) I-II軸成分



(b) I-III軸成分

図6 数量化理論第Ⅲ類の散布図

“回” “団” “国” “園” の例

を算出し、識別平面における第Ⅰ軸のスコアとする。次に“団”，“園”の字種の外的基準を+1，“回”，“国”の外的基準を-1として、先程と同じように学習文字と未知文字に対する予測値を算出し、識別平面における第Ⅱ軸のスコアとする。つまり図7のような外的基準を設けて、各字種のスコアを算出することになる。このスコアを散布図上にプロットし、この散布図上の軸を境界線として識別操作を行う。

ところがこのようにしてスコアを求めると、特徴項目数が学習サンプル数より大きくなるため、数量化理論第Ⅰ類の分析結果は意味を持たないものになってしまう。一般に数量化理論第Ⅰ類の分析を行うための必要条件としてサンプル数を n 、特徴項目数を R としたとき

$$n \geq R + 1 \quad (9)$$

が成り立つ必要がある。したがって、分析結果が意味のあるものにするためには、サンプルの数を増やすか特徴項目の数を減らす操作を行わなければならない。サンプルの数を増やすことは容易ではないので、特徴項目の数を減らすことを考える。この操作には先程述べた相関比による特徴項目選択を用いる。

ここでは選択する特徴項目数と、数量化理論第Ⅰ類により識別特徴抽出したときの識別率との関係を検討する。類似文字集合No.1のデータを用いた例を図8に示す。図8は識別率を縦軸に、特徴項目数を横軸に示したグラフである。学習文字の場合は項目数を80項目まで減らしても100%の識別率を保っている。しかしこれ以上項目数を減らして行くと識別率が下降していく。これは、3章で行った距離による識別の結果と同様、項目数を極端に減少させると、手書き文字から抽出した文字特徴の変動に対処できなくなるためだと考えられる。ところが未知文字の識別結果を見ると、特徴項目数が多くても識別率が良いとは限らないことが分かる。未知文字の場合には特徴項目数を減らして行くと識別率が上昇し、特徴項目数200項目で最も高い識

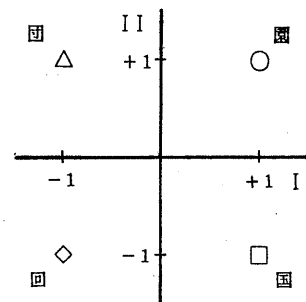


図7 数量化理論第Ⅲ類の分析結果より決定した外的基準
“回” “団” “国” “園” の例

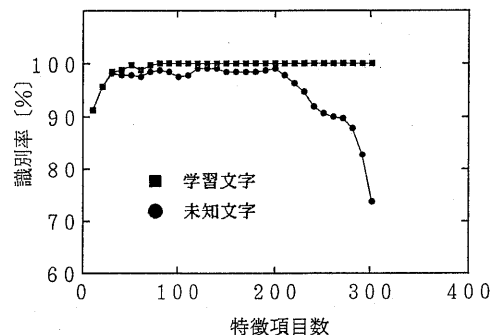


図8 数量化理論第Ⅰ類による識別結果
“回” “団” “国” “園” の例

表6 識別結果の比較

類似文字集合	数量化I類による最大識別率と 同項目数の距離による識別率			距離による最大識別率		距離による 識別率 全特徴項目使用
	項目数	数量化I類	距離	項目数	識別率	
No.1	120項目	99.1%	99.1%	90項目	99.1%	99.1%
No.2	80項目	96.3%	92.8%	145項目	93.4%	92.2%
No.3	80項目	96.3%	91.9%	300項目	94.1%	93.8%
No.4	80項目	95.9%	92.8%	365項目	97.8%	97.5%

別率99.1%が得られている。また120項目で、99.1%が得られる。このときの未知文字の散布図を図9に示す。項目数をさらに減らすと学習文字の識別率が大きく低下すると共に未知文字の識別率も大きく低下していくことが分かる。この低下は学習文字での識別率の低下と同じように、項目数を極端に減少させると手書き文字から抽出した文字特徴の変動に対処できなくなり、識別率が低下するものと考えられる。

使用する特徴項目数が多過ぎる場合も、未知文字の識別率が低下する。特徴項目数がサンプル数に近づく、散布図上における学習文字のプロットのばらつきは抑えられる。しかし未知文字の写像は意図するところに行われず、結果として識別率が低下する。

数量化理論第I類の立場からみると、サンプル数は特徴項目数に比べて十分大きくとる必要がある。ところが文字の識別の精度を上げるとい立場からみると、特徴項目数を減らし過ぎることはできない。用いることができる学習サンプルの数に限りがある場合もあるので、これら相互の作用で最も良い識別率を得られる特徴項目数を探す必要がある。

4.4 距離による識別結果との比較

数量化理論第I類による識別結果と距離による識別結果を表6に示す。類似文字集合No.2、No.3については数量化理論第I類の識別結果が、距離による識別結果を上回っている。No.1については識別率は変わらなかった。No.4については距離による識別の方が数量化理論第I類より良かったが、特徴項目数を同じ所で考えると数量化理論第I類の方が識別率はよい。したがって、No.1、No.4についてもサンプル数を増すことによって識別率の向上が期待される。

4.5 外的基準に関する検討

数量化理論第I類による識別では、平面散布図上の各象限にカテゴリ配置し、平面散布図上の2方向の軸を識別境界線とすることで、カテゴリの識別を行った。また、それぞれの文字の配置は、数量化理論第III類の分析結果の位置関係をもとに決定した。ところがこのように各象限に4つのカテゴリを配置する場合図10に示すように、全部で3通りの配置が考えられる。数量化理論第III類の分析結果をもとに外的基準を決定したものを外的基準1、その他2つの外的基準を外的基準2、外的基準3とする。この3通りの外的基準をそれぞれ用いて、数量化理論第I類により識別を行った。その結果を図11に示す。図11では、ほとんどの部分において、外的基準1の識別率が外的基準2、外的基準3の識別率を上回っている。数量化理論第III類は、

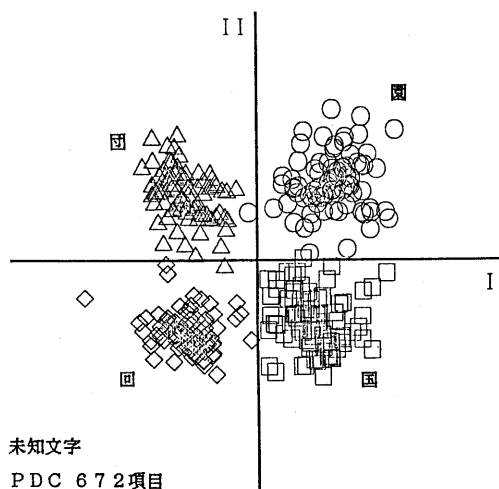


図9 数量化理論第I類による散布図
“回” “団” “国” “園” の例

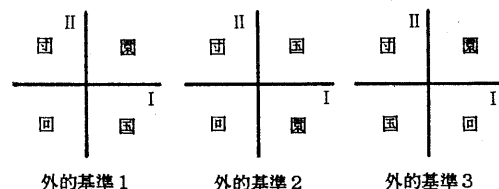


図10 外的基準の与え方
“回” “団” “国” “園” の例

表7 3次元識別空間による識別結果

類似文字集合	数量化I類による最大識別率と 同項目数の距離による識別率				距離による最大識別率	
	項目数	3次元	2次元	距離	項目数	識別率
No.1	80項目	100 %	99.1 %	99.1 %	90項目	99.1 %
No.4	80項目	97.2 %	95.9 %	92.8 %	365項目	97.8 %

もとの高次元特徴空間のカテゴリの分布特性を低次元空間に写像する。数量化理論第I類の外的基準も、もとの構造特徴空間のカテゴリの分布を反映させることにより、識別率が向上できると考えられる。したがって、外的基準を決定するために数量化理論第III類の結果を用いることが有効であることが分かる。

4.6 4章まとめ

数量化理論第III類の分析結果をもとに外的基準を決定することにより数量化理論第I類の識別率を向上できることが分かった。また、数量化理論第I類の未知文字に対する識別率は用いる特徴項目数を適当に決めることによって向上できることが分かった。

5. 3次元配置による識別

4章では平面散布図上で識別の操作を行った。この際外的基準の与え方は3種類あった。そこでは3種類の外的基準から2つを選び平面散布図上の2つの軸にあてはめた。しかし識別空間を3次元にすることにより3種類の外的基準を同時に取り入れることができる。この場合識別空間上のサンプルは外的基準からのユークリッド2乗距離で識別する。識別空間の次元数を増すことにより、識別率の向上が期待できる。そこで、4章で識別率が向上しなかった類似文字集合No.1とNo.4についてこの手法を用いて詳細識別を行った。この結果を表7に示す。どちらの例も2次元で識別を行ったときよりも識別率が向上している。しかし類似文字集合No.4では距離による識別の結果を上回ることはできなかった。

6. まとめ

数量化理論第I類を用いて類似手書き文字の詳細識別を試みた。その結果数量化理論第I類では用いる特徴項目数によって識別率が変化することが分かった。特徴項目数を適当な値に設定することにより、低次元の識別空間でも高い識別率が得られることが分かった。

謝辞 実験に使用した手書き教育漢字データベースETL8Bの利用に関してご配慮下さった電子総合研究所の山田博三氏、他 関係各位に深謝する。また計算機利用にご配慮下さった本学習志野計算機センターの斉藤技手に感謝する。最後に日頃御討論下さる井上講師、大学院の熊野君、実験を担当してくれた学部4年の木島君並びに研究に協力してくれる学部4年の小八重君、他 研究室諸氏に感謝する。

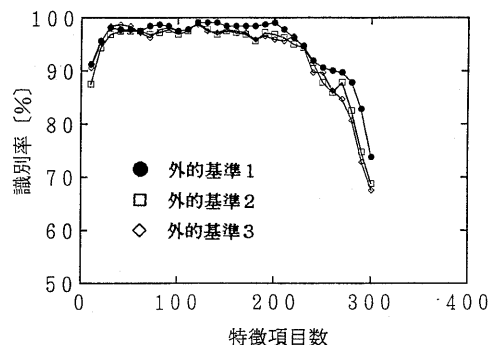


図11 外的基準の変更による数量化理論第I類の識別結果
“回” “団” “国” “園”の未知文字の例

参考文献

- 1) 坂井邦夫: “文字文書認識と理解”, 信学誌, 71, 11(1988).
- 2) 例えば 飯島泰蔵: “図形の特徴抽出に関する基礎理論”, 信学論(c), 54-C, 12(1971).
- 3) 森 他: “手書き教育漢字のデータベースについて”, 電総研彙報, 43, 11-12(1979).
- 4) 斉藤他: “手書き文字データベースの解析(V)”, 電総研彙報, 45, 1-2(1981).
- 5) 萩田他: “外郭方向寄与度特徴による手書き漢字の識別”, 信学論(D), J66-D, 10(1983).
- 6) 太田他: “相関比を用いた文字の特徴項目の選択に関する一検討”, 信学秋全大, D-206(1988).
- 7) 田中他: “多変量統計解析法”, 現代数学社(1983).
- 8) 西脇他: “数量化理論第III類による類似文字の詳細分類”, 信学論(D), J69-D, 12(1986).
- 9) 西脇他: “数量化理論第III類による文字の特徴選択”, 信学論(D), J70-D, 12(1987).
- 10) 林知己夫: “数量化の方法”, 東洋経済新報社(1982).