漢字情報を用いた筆跡の矩形診断に関する研究

2003MM073　中村元樹
指導教員: 松田眞一

1　はじめに

古くから、個人を特定するものの一つとして筆跡による鑑定が扱われてきた。しかし、近年の急激なＩＴの進歩により文字を手で書くということが少なくなってきている。それでも、遺書などは手で書かれており、大事な書類などには署名を使っている。そこで、筆跡でどれだけ人を判別できるのかを知りたくなったのでこの研究をしようと思った。

また、日本には、ひらがな、カタカナ、漢字と多くの文字が存在する。この研究では、その中で漢字に注目し、その文字の外接を長方形で囲むことで計数化する矩形診断法を使い、さらに、ペン字の見本の載っている辞書[1]からサンプルの文字を使って、その漢字情報との比較から鑑定しようと試みた。

2　データについて

今回扱うデータは漢字に注目するため、四つの漢字で構成されている四字熟語を用いることにした。その中で、偏や旁などの部首や形がなるべく均等になるように、「異口同音」「自然淘汰」「国士無双」「免許皆伝」「新陳代謝」の5つを選び、この5つの四字熟語について10回ずつアンケート実施者に書いてもらった。そして、書いてもらった文字をスキャナでコンピュータに400dpiの画素数で読み込み、“Windows”に標準装備されている“Paint”を使ってひとつひとつ外接長方形を作り、そのドット数を長さとして用いた。

また、紙面上の都合により、「異口同音(異口)」、「自然淘汰(自然)」、「国士無双(国士)」、「免許皆伝(免許)」、「新陳代謝(新陳)」とする。

3　分析方法

分析方法は、判別分析と交差確認法(cross validation)を用い、古橋・長谷川・伊藤・浦末[2]を参考に、図1のように、矩形診断法で置き換えられた外接長方形から「縦の長さ(A)」、「横の長さ(B)」、「アンダーラインから底辺までの高さ(C)」、「間隔(D)」、「『縦』/『横』の比(E)」「漢字情報との比較(R)」を変数として扱った。

また、(R)は、

$$R = \frac{B}{\frac{A}{CA} \times (CB)} \tag{1}$$

の計算から導いたものを用いた。(ここで、CAとは漢字情報のたて、CBとは漢字情報のよこを表すものとする。)

4　分析結果

今回の分析では、50人のアンケート実施者から得たサンプルから、今回は線形判別分析法だけを用いて「交差確認法を使わない場合」、「leave-one-out cross validation」、「10分割交差確認法(10-fold cross validation)」、「2分割交差確認法(2-fold cross validation)」の4種類に対して、漢字情報を追加する前の判別率と漢字情報を用いた場合の判別率を比較することにした。

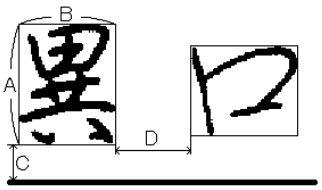

図 1: 変数の説明

判別率の算出方法はMingzhe Jinのホームページ「Rと判別分析」[3]で使われている方法を用いた。また、漢字情報を用いた場合で他の変数を減らした場合と漢字情報を追加する前とを比較した。

4.1　記号の定義

ここに表で用いる記号を下のように定義する。

(ア):「A」「B」「C」「D」を用いた場合

(イ):(ア)から「A」「B」を除いて「E」を加えた場合用いた場合

(ウ):「A」「B」「C」「D」「R」を用いた場合

(エ):(ウ)から「A」を除いた場合

(オ):(ウ)から「B」を除いた場合

(カ):(ウ)から「A」「B」を除いた場合

「交差確認法を使わない場合」ではトレーニングデータとテストデータを同じもので判別しているため判別率を過大評価している。そのため非常に高い判別率が得られるため、ここでは交差確認法を使った場合だけを載せる。

4.2　leave-one-out cross validation

まず、「Leave-one-out cross validation」を使って判別した。

表 1: leave-one-out cross validation(%)

	漢字情報なし		漢字情報あり			
	ア	イ	ウ	エ	オ	カ
異口	90.4	72.2	90.4	89.0	89.6	73.6
自然	90.0	72.0	89.6	89.4	89.0	72.8
国士	90.2	70.0	89.0	90.8	89.6	73.0
免許	88.2	70.4	86.8	89.0	87.8	74.4
新陳	86.4	70.8	87.4	86.8	86.4	75.2

「漢字情報なし」と「漢字情報あり」で比較してみると

「自然淘汰」以外では(ウ)(エ)(オ)のいずれかが高いか変わらない判別率が得られた。また、(イ)と(カ)を比較してみるとどの場合でも(カ)の方が高い判別率となっている。

4.3 10分割交差確認法(10-fold cross validation)

各群の10個のデータを10個に分割して交差確認する「10分割交差確認法」を行った。結果は表1と比べると(イ)や(カ)では10%ほど、その他では5%ほど判別率が上がった。また、詳しく見てみると後半に書かれた文字に比べて前半に書かれた文字は判別率が低いことが分かった。

4.4 2分割交差確認法(2-fold cross validation)

各群のデータを前半5つ後半5つに分割して交差確認する「2分割交差確認法」を行った。また、奇数番目に書かれた文字と偶数番目に書かれた文字に分割しても行った。結果はどちらの場合も表1と比べて10%ほど低く若干前半と後半で分割した方が良い判別率となった。

4.5 考察

文字の直接な長さである「たて」と「よこ」を除いた場合の(イ)と(カ)を比較してみるとどの方法でも(カ)の場合の方が高く、そういった意味においては漢字情報は有効であり、このことから文字を書く用紙によって出る文字の大きさの違いに対して有効であると思われる。次に(ア)と(ウ)(エ)(オ)を比較してみるとほとんどの場合で漢字情報を用いた場合のいずれかで高い判別率が得られている。また、(エ)と(オ)を比較してみると前後半で分割した2分割交差確認法以外のどの方法でも判別率の高くなっているものは同じで、四字熟語によって影響が強いのは「たて」か「よこ」かはっきりと分かれていると感じられた。唯一異なっていた前後半で分割した方法ではもともと少ないトレーニングデータが半分の5つになってしまったためと見かけの安定性がなくなったためと思われる。

5 4つの漢字をまとめた場合の判別

本研究では漢字情報を扱っているため四字熟語の異なった4つの漢字を同じものと考えて判別した。「たて」「高さ」「漢字情報」の3つを変数として「Leave-one-out cross validation」と「2分割交差確認法」を行った。

5.1 結果と考察

どちらの方法でも15%〜25%ほどの非常に低い判別率となった。また2分割交差確認法では前半→後半の判別率に比べて後半→前半の判別率が非常に低かったため、ここでも奇数と偶数で分割して行い、その結果あまり差のない判別率が得られた。判別率の低い原因としてやはりもともと違う文字を同じものとしているためであると考えられるが、それでも2割は判別できることが分かった。ただし、どのように判別されているか詳しく見てみると、27番や40番が多く正確に判別されていることがわかり、その理由として他の文字の大きさと比べて27番は極端に小さく40番は極端に大きいためであり、どの人の文字も2割程度判別されているわけではないことがわかる。また、この方法からも前後半で分割した場合の不安定さが感じられる。

6 各漢字の平均を用いた場合の判別

次に、四字熟語の各漢字の「たて」や「よこ」の変数を平均して一つの変数としたときに判別率がどうなるか考えた。ここでは「Leave-one-out cross validation」を扱った判別分析を行った。

6.1 結果と考察

「たて」「よこ」「高さ」「間隔」「漢字情報」の全ての平均を用いても70%近い判別率が得られた。また、最も高い判別率が得られたのは「漢字情報」だけを平均にした場合で、「国士無双」「免許皆伝」を見てみると、4.2の(ウ)よりも高く、「自然淘汰」ではどの場合の判別率よりも高くなった。このことから、平均を用いることは扱う変数も少なくなり、十分に有効であると考えられる。

7 まとめ

先輩の行った筆跡鑑定の方法に漢字情報を加えても直接判別率に大きな影響は与えなかった。しかし、漢字情報がない場合では「たて」や「よこ」の変数を除くと判別率を下げてしまうのに対して漢字情報がある場合はさらに精度の高いものとなることが分かり、異なった文字の大きさで判別する際でも有効であることが分かった。そういった意味では、漢字情報という変数を追加したことに大きな意味があったといえる。さらに、この変数を追加したことでさまざまな方法が考えられるようになった。しかし、今回扱った方法ではあまり良い結果は得られなかったため残念である。

8 おわりに

本研究を行って、筆跡には個人の特徴が良く現れるものであることは前々から分かってはいたが、今回のように、中の情報は一切使わずにただ文字の外接長方形をとっただけで判別してもかなりの確率で判別できることに驚いた。この方法にさらに中の情報を混ぜたらかなりよい結果が得られるのだろうと思う。そして、今回は50人のデータで判別したが14人で判別したときとほぼ変わらない判別率が得られたことに驚いた。また、この研究を行うにあたり、計算に関しては統計ソフト「R」を使えば楽にできるが、データを取るために50人分の四字熟語を長い時間をかけて数値にしていく作業はかなりつらく、統計解析の大変さを痛感した。さらに、「判別分析」は大学の授業ではあまり勉強していなかったため、「R」の使い方や理論など分からないことだらけであったが、この機会に勉強することができ大変良かったと思っている。

参考文献

[1] 石井庄司: 大きな字の常用国語辞典 改訂第三版, 学習研究社,2006.

[2] 古橋あい・長谷川千津・伊藤志麻・浦末直樹: 統計的解析による筆跡鑑定, 南山大学経営学部情報管理学科卒業論文要旨集,1996.

[3] Mingzhe Jin: Rと判別分析, http://www1.doshisha.ac.jp/~mjin/R/17.html.