1F-08

輪郭線方向成分とZipf則を用いた文字画像の自動分類

山口 文彦
長崎県立大学 教育開発センター

1 はじめに

近年，自然言語処理の手法を失われた言語に適用する研究が盛んである [1]。こうした研究の多くは，対象とする言語のテキスト情報を元に，さまざまな処理を行なう。テキスト情報は文字の並びであるが，これを得るには，木片・粘土板・石板などの遺物の写真や画像から，文字と思われる記号を切り出し，同じ文字の異なる出現を同定する必要がある。従来この過程は，考古学者の経験に基いて，人の手で行われてきた。しかし，記号が多いときには多人数で作業することもあり，作業者の熟練度合いなどによって遺物の画像に現れる記号を文字へと分類する際に判断が揺れることがある。また，未解読の文字の場合，正解が設定できないことから，恣意的な分類になってしまう危険もある。

こうした問題に対処するため，山口は記号の画像から文字クラスを自動的に抽出する研究を行なっている [2, 3]。これらの研究で対象としているのはイースター島で製作された木片に刻まれたロンゴロンゴと呼ばれる記号の列だが，ロンゴロンゴは未解読であるため，結果の評価が難しい。そこで本稿では，既知言語である日本語を対象に，画像から文字クラスの抽出を試み，提案手法の評価と問題点の指摘を行なう。

2 記号画像から文字クラスへの分類

記号ごとに切り出された画像情報を要素とする集合に対し，この集合を同じ文字を表す記号の集合に分割する問題を考える。

同じ文字であれば，その表現である記号の画像は似た特徴を持つと考えられる。しかし，異なる文字が似ていることもあるので，画像特徴量の近さだけで文字へと分類することは難しいと考えられる。そこで，まず画像特徴量を基に記号画像の集合を階層的に分類し，そののち，文字の出現頻度が Zipf 則 [4] に従うと仮定して階層的な分類を統合することで，非階層的な分類を得る方法を提案する。

画像特徴量として，手書き文字認識 [5] でよく用いられる輪郭線方向成分と記号領域の縦横比を用いた。階層的分類の手法としては，類似度の比較にコサイン距離を用い，クラスタの構成に UPGMA 法を用いた。階層的な分類を統合して階層的な分類を得るには，Zipf 則に従う度合いを評価値とする遺伝的アルゴリズムを用いた。

2.1 非階層的分類の Zipf 則による評価

Zipf 則は，頻度が k 番目に大きい要素が全体に占める割合が，ある定数 s があって $1/k^s$ に比例するという経験則である。非階層的分類に対して，その各クラスタに含まれる要素数と，クラスタの大きさの順位を求めることができる。Zipf 側は，これが両対数のグラフ上で直線に近似できることを言っているので，最小二乗法によって近似直線を求め，この近似直線との誤差を求めることができる。また，Zipf 則は直線の y-切片も規定するので，近似直線の y-切片との差も求めることができる。これらが小さいほど，得られた非階層的分類が Zipf 則に従っていると言える。

2.2 遺伝的アルゴリズムによる解探索

上記の評価値を用いて，非階層的分類を遺伝的アルゴリズムによって求める。ここでは，非階層的分類の仕方が解である。解の表現は，階層的分類で得られる二分木の，根を含む部分木である。したがって，解表現も二分木となる。この解表現である部分木の葉のそれぞれが，非階層的クラスタのそれぞれを表す。

解表現を掛け合せる cross-over では，部分的な構造を継承するだけでなく，得られた表現が解を表していなければならない。そこで，二つの解表現 (親となる二分木) の根から同じ位置にある節点を一つランダムに選び，これらを交換する方法を用いる。こうすることで，cross-over によって生成された解表現 (子となる

Automated Classification of Character Image using Outline and Zipf's Law
Fumihiko YAMAGUCHI
Education Development Center, University of Nagasaki

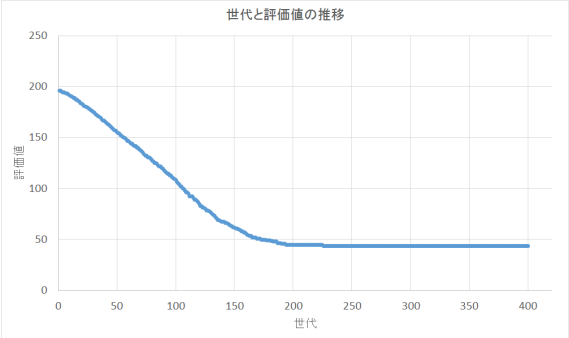

図 1: 世代ごとの評価値の推移

二分木) も解 (非階層的分類) を確実に表現し，また解の構造も部分的に継承される。

3 実験と結果

本研究では，上記の手法の有効性について検証するため，日本語の文字を対象とする。日本語の文字データとして，東京農工大学中川研究室で配布しているオンライン手書き文字パタンデータベース [6] TUAT Nakagawa Lab. kondate-14-09-01 を用いた。本研究では筆順を考慮しないオフライン文字パタンを用いるので，kondate-14-09-01 の点を結んだ画像を用いた。また，文字の出現頻度が重要な特徴となるので，このデータベース中で被験者が自由に記述した項目を対象としている。

提案手法によって得られた記号の分類が，正しく文字としての分類になっているかを評価するために，各文字についてその文字を含むクラスタの個数と，得られた各クラスタに含まれる文字の種類数を調べた。いずれも 1 以上の値をとり，1 であることが望ましい。

GA を 400 世代実行した際の，各世代における最小評価値の推移を図 1 に示す。200 世代以降ほとんど変化していないことから，200 世代前後で収束している様子が分かる。400 世代経過後の最小評価値となる非階層的分類において，各文字を含むクラスタの個数は平均して 1.90 個，得られた各クラスタに含まれる文字の異なり数は平均して 9.38 個となった。

そもそも Zipf 則を用いて得られる評価値が，正しく文字を分類できるかという問題がある。そこで個体の評価値と結果の評価値の関係を調べた。いずれも小さいほど良いので，正の相関があることが望ましい。しかし，図 2 に示した個体の評価値と結果の評価から，各クラスタが含む文字の異なり数は，Zipf 則から得られる個体の評価値とは (相関があるとしても) 負の相関となっていることが分かる。

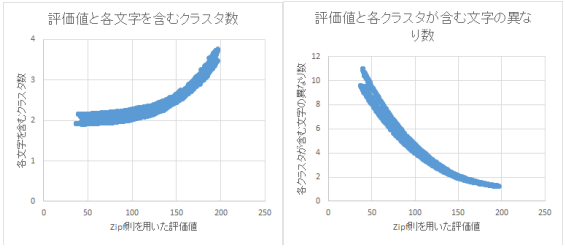

図 2: 個体の評価値と結果の評価

4 結論

日本語の文字画像に対し，画像特徴量と Zipf 則を用いて，文字クラスを自動的に得ようとする実験を行なった。結果として，同じ文字が少数のクラスタに属する結果が得られたが，各クラスタが複数の文字を含んでしまう結果となった。

今後，個体評価値に画像特徴量を含めるなど，同じクラスタに異なる文字が含まれないようにする工夫が必要である。

謝辞

本研究は MEXT 科研費 24500313 の助成を受けたものです。

参考文献

[1] B.Snyder, R.Barzilay, and K.Knight, “A Statistical Model for Lost Language Decipherment”, ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 1048–1057, 2010

[2] 山口文彦, “Zipf 則を指標とするロンゴロンゴ記号の分類”, 日本情報考古学会講演論文集, vol.15(通巻 35 号), 研究発表 17 , 2015

[3] 山口文彦, “手書き文字認識手法を用いたロンゴロンゴ記号の類似度”, 日本情報考古学会講演論文集, vol.13(通巻 33 号), 研究発表 19 , 2014

[4] David M.W. Powers, “Applications and explanations of Zipf’s Law, New Methods in Language Processing and Computational Natural Language Learning”, ACL, pp.151–160, 1998

[5] 鶴岡信治, 栗田昌徳, 原田智夫, 木村文隆, 三宅康二, “加重方向指数ヒストグラム法による手書き漢字・ひらがな認識”, 信学論 (D), J70-D, no.7, pp.1390–1397, 1987

[6] M.Nakagawa, K.Matsumoto, “Collection of on-line handwritten Japanese character pattern databases and their analysis,” Int. J. Doc. Anal. Recognit., vol. 7, no. 1, pp. 69–81, 2004