単語の長さの情報に基づいた文章のパターンの分類*

5L－10

金明哲†
総合研究大学院大学統計科学‡

概要

著者不明の文章の著者の推定などを行なう際、文章の中のどんな情報を用いるかが鍵である。著者の個性が十分含まれていない情報では、有効な方法でも著者の推定、著者別の文章の分類が正しくできないのは言うまでもない。欧米における文章については長年にかけ多くの研究が行なっているが、日本文の場合は著者の推定などに用いる情報に関しては研究が十分行なっていない。

本文では品詞別にわけていない場合と品詞別に分けた場合の単語の長さの情報に基づいて著者別の文章の分類について分析を行なった。分類がもっとも良いのは動詞の長さの情報で、品詞別に分けていない、すべての単語の長さの情報を用いた場合よりはるかに良い。

分析の手法としては群内距離と最小の群間の距離との差の大小を統計量・分散比で評価し、分散比が大きいほど分類が良いと判断した。分類結果を視覚化する手法としては主成分分析を用いた。

1　はじめに

欧米における文献の著者の推定や文体の研究では、文の長さ、単語の長さ、単語の使用頻度などに関する情報がよく用いられている。しかし、日本文は英文のように"分ち書き"されていないため、単語に関する情報はあまり分析されていない。

文章の著者の推定などを行う場合、文章のどのような情報に注目すべきかは言語の種類によって異なると考えられる。例えば、日本語では漢字と仮名が混じっているため漢字の使用率に著者の特徴が現れる可能性がある [1]。また、日本語では読点の付け方に明確な規則がないため、読点の情報も著者の特徴になる [2][3]。では、単語の長さはどうであろうか。残念ながら日本語の場合、単語の長さの情報に関する基礎的な研究はない。そこで、本研究では単語の長さと著者の関係を明らかにするため、井上靖、三島由起夫、中島敦3人の21編の文章について、単語の長さの情報を用いて著者別に文章の分類を行なう場合、それぞれの作家のものとして、うまく 分類されるかを試みた。

2　群内と群間での距離による分類分析

分類の立場からは用いた情報に基づいた群内の任意の二つの分布間の距離は群間での任意の二つの分布間の距離より小さいことが望ましい。そこで、群内での任意の二つの分布の間の平均距離と群間での任意の二つの分布の間の平均距離を求め比較を行なった。

用いた距離は次のように求める。いま、作品 i での長さ j の単語の出現率を p_{ij} と表すと、作品 i における長さ J までの出現率は

$$p_{i1}, p_{i2}, \cdots, p_{ij}, \cdots, p_{iJ}$$

に表示できる。同じに、作品 l については

$$p_{l1}, p_{l2}, \cdots, p_{lj}, \cdots, p_{lJ}$$

に表示できる。この二つのの分布の間の距離 $d(i,l)$ を次のように定義する。

$$d(i,l) = \frac{1}{2}\sum_{j=1}^{J}(p_{ij}log\frac{2p_{ij}}{p_{ij}+p_{lj}} + p_{lj}log\frac{2p_{lj}}{p_{ij}+p_{lj}})$$

3人の21文章について、品詞別に分けていない場合 (すべての単語) と品詞別に分けた場合の群内、群間での任意の二つの分布間の距離の平均値を求めた。群内の距離と群間の距離を比較してみると、群内での距離が群間での距離より小さいのもあれば、群内での距離が群間での距離より大きいのもある。すべての単語と品詞別に分けた場合で、どの情報を用いた方が分類がもっと良いであろうか、これは我々が興味を持っている問題である。分類を良くするためには、最小の群間の距離が群内の距離より大きいければ大きいほどよいと考えるため、群内の距離と最小の群間の距離との差で分類の良さを議論すれば良い。

いま、著者 i での群内での距離を y_{i1}、最小の群間の距離を y_{i2} に表すと I 人の著者の群内の距離、最小の群間の

*Classifiction of Styles by Information of Word Length

†Ming-Zhe Jin ,Email:Jin@iam.ac.jp

‡Department of Statistic1 Science, The Graduate University for Advanced Studies,4-6-4 Minamiazabu,Minatoku,Tokyo,Japan

距離はマトリックス $Y_{I\times 2}$ に表示できる。

$$Y_{I\times 2} = \left[\begin{array}{cc} y_{i1} & y_{i2} \end{array} \right]$$

$Y_{I\times 2}$ の総変動は

$$S = SS_A + SS_B + SS_E$$

に分解することができる。式の右辺はそれぞれ行の変動、列の変動、誤差の変動である。それぞれの自由度で基準化したのを

$$S_A^2 = SS_A/(I-1), S_B^2 = SS_B, S_E^2 = SS_E/(I-1)$$

に表す。行間の差、列間の差はそれぞれ $S_A^2/S_E^2, S_B^2/S_E^2$ の分散比で評価すれば良い。分散比の値が大きければ大きいほど行間、列間の差が大きいことを説明する。
我々が興味をもっているのは群内の距離と最小の群間の距離との差の大きさで、分散比 S_B^2/S_E^2 が大きいのが分類がよいと判断する。すべての単語を用いた場合と品詞別に分けた場合の分散比 S_B^2/S_E^2 を求め表 1 に示した。

表 1: 群内の距離と最小の群間の距離との分散比

	すべての単語	名詞	動詞	形容詞
S_B^2/S_E^2	2.0555	1.4668	46.2447	0.4185

	形動	助詞	助動詞	副詞
S_B^2/S_E^2	22.3156	0.9811	0.8100	1.4573

分散比 S_B^2/S_E^2 がもっとも大きいのは動詞で、その次は形容動詞である。したがって、動詞の長さの分布の情報を用いた場合、著者別の文章の分類がもっとも良いと判断する。

3 分類のプロット

分類の視覚化にはいくつの方法があるが、本文では主成分分析を用いた。作品 i の長さ j の出現率を p_{ij} に表すと、I 編の作品の長さ J までの出現率の分布はマトリックス

$$P_{I\times J} = [p_{ij}]$$

に表記できる。主成分分析は $P_{I\times J}$ の分散共分散行列を用いて行なった。分散比が最も大きい動詞の長さの情報を用いた場合の、第 2 主成分までの主成分得点のプロットを図 1 に示した。第 2 主成分までは累積寄与率は 83.77% である。品詞別に分けなかった場合と比較を行なうため、すべての単語の長さの情報を用いた場合の主成分得点のプロットを図 2 に示した。第 2 主成分までの累積寄与率は 97.92% である。動詞の長さの情報を用いた場合の累積寄与率はすべての単語の長さの情報を用いた場合より低いにもかかわらず、分類は動詞の長さの情報を用いた場合が良い。

図 1: 動詞の長さの情報の主成分のプロット

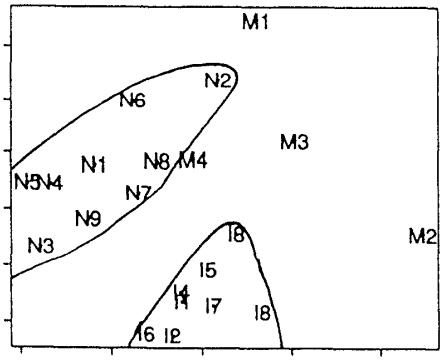
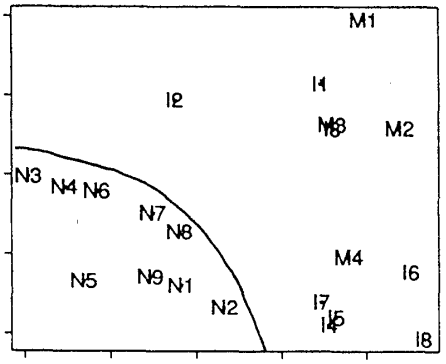


図 2: すべての単語の長さの情報の主成分のプロット



4 おわりに

分析に用いた文章では、動詞の長さの情報に基づいた分類がすべての単語の長さの情報を用いた場合より分類が良いことが分かった。その普遍性について研究を続ける必要がある。

多額の研究費でデータベースを作成し、本研究を支援して下さった文部省統計数理研究所村上征勝様に深く感謝致します。

参考文献

[1] 金 明哲、樺島 忠夫、村上 征勝 (1993c). 手書きとワープロによる文章の計量分析、計量国語学, Vol.19, 133-145.

[2] M.Z.Jin and M.Murakami(1993a). *Authors' Characterristic Writing Styles as Seen Through Their Use of Commas,* Behaviormetrika, Vol.20, 63-76.

[3] 金 明哲、樺島 忠夫、村上 征勝 (1993b). 読点と書き手の個性、計量国語学、Vol.18, 382-391.