助詞の分布における書き手の特徴に関する計量分析

金 明哲

In the study, statistical analyses about distribution of particles were done by the method of distance analysis, analysis of variance, classification analysis, discriminant analysis and discrimination rules mining. As a result, it was clarified that the distribution of particles differed in each writer and it is a powerful characteristic when recognizing the writers of texts.

Keywords: Recognizing the Writer, Distribution of Particles, Distance of Distribution, Classification Analysis, Discriminant Analysis, Discrimination Rules

キーワード：書き手の識別，助詞の分布，分布間の距離，分類分析，判別分析，識別ルール

1．はじめに

文章（書）から書き手の文体の計量的な特徴を抽出し，その統計分析によって文章の書き手を識別（認識，判別，推定）する研究では，書き手の特徴情報として文章に関するどのような要素を用いるべきであるかが問題解決の鍵である．文章の書き手の識別などの研究では，欧米文では文の長さ，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（Holmes 1994）．日本文に関しては，文の長さの分布（安本 1957 a，1957 b，1994，佐々木 1976），品詞の使用率（村上 1994）が書き手の特徴情報としてよく用いられている．これらの特徴情報に関しては，書き手の特徴を表す情報の一つであることは否定しないが，しばしば書き手の特徴が明確に現れないことが実証されている（金 1994 a）．特定の単語に関しては，安本は直喩，声喩，色彩語，人格語などの使用頻度を用いて源氏物語の書き手の推定を試み（安本 1994），韮沢は「にて」，「へ」，「して」，「ど」，「ばかり」，「しも」，「のみ」，「ころ」，「なむ」，「じ」，「ざる」，「つ」，「む」，「あるは」，「されど」，「しかれども」，「いと」，「いかに」などの単語を用いて「由良物語」の書き手の判別を試みた（韮沢 1965）．前者の直喩，声喩，色彩語，人格語などは文章のなかに占める割合が低いため，安定した情報が得られない可能性があり，後者は特定の書き手の特徴について研究を重ね選定した単語であるため，一般性が欠けている．

不特定の書き手の特徴情報としては，多くの人が文章を書くとき頻繁に使う単語（要素）であることが望ましい．このような条件

JIN Mingzhe 札幌学院大学社会情報学部

を満たし，かつ書き手の特徴が明確に現れる単語は何であろうか？ 日本文に関しては，このような基礎的な研究が十分とはいいがたい．文章のどのような要素に書き手の特徴が現れるやすいかに関しては，言語の種類によって異なる．そこで，筆者は日本現代文における書き手の特徴情報の抽出に関する研究に取り組み，いくつかの研究成果をあげている（Jin and Murakami 1993，金・他 1993 a，1993 b，金 1994 a～1996 d）．

本稿では，その一環として 3 人の 28 の文章（合計 126844 単語）を用いて，品詞別に見た場合，使用頻度が最も高い助詞（約 30%～40%）の分布に書き手の特徴が現れるか否かについて分析を行うと同時に，データマイニング法による書き手の識別ルールの抽出に関する試みについて述べる．

2．分析に用いた文章

統計分析に用いたのは井上靖，三島由紀夫，中島敦の短篇小説である．分析に用いたデータベース作成の労力の節約および文章から抽出された情報の安定性を見るため，これらの中で比較的長い文章は短い文章にあわせ，いくつかに分割した．例えば，井上の「恋と死と波と」は二つに，三島の「潮騒」は七つに，中島の「弟子」は三つに，「李陵」は四つに分割して用いることにした．表 1 に，用いた文章と発表年やサイズなどを示した．単語の認定基準としては「長い単位」を用いた．品詞の認定は『広辞苑』，『国語辞典』（旺文社，第八版）に基づいた．ただし，複合語は 1 語と見なした．

表 1．分析に用いた文章のリスト

書き手	文章名	文章の記号	単語数	出版社	発表の年
井上 靖	結婚記念日	I 1	4749	角川文庫	1951
	石庭	I 2	4796	同上	1950
	死と恋と波と	I 3	4683	同上	1950
	死と恋と波と	I 4	4386	同上	同上
	帽子	I 5	3724	新潮文庫	1973
	魔法壜	I 6	3624	同上	同上
	滝へ降りる道	I 7	3727	同上	1952
	晩夏	I 8	4269	同上	同上
三島 由紀夫	遠乗会	M 1	4984	新潮文庫	1951
	卵	M 2	4004	同上	1955
	詩を書く少年	M 3	4502	同上	1955
	海と夕焼	M 4	3359	同上	1955
	潮騒 1	M 5	5556	同上	1955
	潮騒 2	M 6	5276	同上	1955
	潮騒 3	M 7	6105	同上	1955
	潮騒 4	M 8	5981	同上	1955
	潮騒 5	M 9	5785	同上	1955
	潮騒 6	M 10	5259	同上	1955
	潮騒 7	M 11	5530	同上	1955
中島 敦	山月記	L 1	3226	新潮文庫	1942
	名人伝	L 2	3202	同上	1942
	弟子 1	L 3	4078	同上	1943
	弟子 2	L 4	4092	同上	同上
	弟子 3	L 5	3727	同上	同上
	李陵 1	L 6	4563	同上	1944
	李陵 2	L 7	4561	同上	同上
	李陵 3	L 8	4638	同上	同上
	李陵 4	L 9	4458	同上	同上

3．品詞の分布

文体の計量分析あるいは文章の書き手の識別などでは，文を構成する要素である単語に関して，どのように計量分析を行うかが分析結果を大きく左右する．一つの文章のなかに現れる，異なる語彙数は文章の長さによって異なる．文章に現れる何百・何千・何万種類の単語をすべて計量分析することは労力がかかるだけではなく，データにノイズを混入させる結果にもなる．文章のなかに現れる単語には文章の内容に大きく依存する単語もあり，内容に関して依存度が低い単語もある．文章の書き手の識別などに用いる情報としては，文章の内容に関して依存度が低く，文章のなかでの出現頻度が高いことが望ましい．どのような単語が文章のなかで頻繁に使用され，かつ文章の内容に関して依存度が低いかに関しても様々なアプローチから研究可能であるが，本研究では品詞別に分けて分析することにする．

表2に10品詞に分けた場合の各文章における品詞の出現率を示した．従来の文体の計量分析には品詞の分布も書き手の特徴情報としてよく用いられているが，表2から分かるように分析に用いた3人のデータでは明らかな差は見られない．このようなデータでは書き手を識別するには無理が伴う（金 1994 a）．品詞の中で出現頻度が最も高いのは助詞で約36%～37%，その次は名詞（約25%～29%），動詞（約16%～18%）の順である．上記の3品詞の出現率が全単語の約80%を占める．本研究では出現率が最も高い助詞について計量分析を行うことにする．

表2．分析に用いた各文章における各品詞の出現率

ID	助詞	名詞	動詞	助動詞	形容詞	副詞	連体詞	形容動詞	接続詞	感動詞
I 1	0.367	0.263	0.174	0.103	0.026	0.036	0.010	0.015	0.007	0.000
I 2	0.350	0.260	0.172	0.119	0.027	0.035	0.013	0.018	0.007	0.000
I 3	0.353	0.261	0.176	0.104	0.027	0.039	0.013	0.019	0.007	0.001
I 4	0.359	0.271	0.174	0.104	0.034	0.028	0.011	0.015	0.005	0.000
I 5	0.366	0.267	0.173	0.125	0.019	0.026	0.010	0.010	0.003	0.000
I 6	0.373	0.252	0.189	0.107	0.018	0.025	0.016	0.015	0.004	0.000
I 7	0.364	0.262	0.178	0.112	0.021	0.031	0.015	0.010	0.006	0.000
I 8	0.370	0.277	0.171	0.100	0.021	0.024	0.017	0.012	0.006	0.001
M 1	0.346	0.280	0.173	0.104	0.021	0.034	0.015	0.023	0.004	0.000
M 2	0.364	0.268	0.176	0.097	0.026	0.029	0.013	0.020	0.007	0.000
M 3	0.365	0.268	0.164	0.100	0.026	0.033	0.012	0.026	0.006	0.000
M 4	0.364	0.276	0.179	0.093	0.025	0.032	0.013	0.011	0.006	0.000
M 5	0.376	0.295	0.172	0.084	0.023	0.025	0.013	0.011	0.002	0.000
M 6	0.363	0.281	0.179	0.106	0.020	0.022	0.013	0.011	0.005	0.000
M 7	0.363	0.281	0.170	0.103	0.024	0.024	0.013	0.018	0.004	0.000
M 8	0.368	0.286	0.169	0.096	0.022	0.024	0.014	0.015	0.005	0.000
M 9	0.365	0.280	0.173	0.099	0.024	0.023	0.014	0.017	0.006	0.000
M10	0.373	0.297	0.165	0.091	0.023	0.024	0.011	0.013	0.003	0.000
M11	0.367	0.290	0.172	0.095	0.021	0.026	0.012	0.013	0.003	0.000
N 1	0.357	0.265	0.164	0.104	0.024	0.046	0.016	0.014	0.008	0.001
N 2	0.365	0.284	0.174	0.091	0.015	0.040	0.014	0.013	0.004	0.000
N 3	0.365	0.259	0.167	0.095	0.027	0.038	0.017	0.025	0.007	0.000
N 4	0.365	0.261	0.169	0.099	0.019	0.041	0.013	0.026	0.006	0.000
N 5	0.364	0.276	0.171	0.093	0.022	0.036	0.016	0.017	0.006	0.001
N 6	0.363	0.294	0.169	0.092	0.023	0.032	0.014	0.012	0.002	0.000
N 7	0.358	0.271	0.167	0.104	0.023	0.034	0.017	0.019	0.005	0.001
N 8	0.357	0.274	0.166	0.107	0.020	0.038	0.014	0.019	0.005	0.000
N 9	0.367	0.269	0.167	0.107	0.017	0.040	0.014	0.015	0.004	0.000

4．計量分析の過程と結果

4．1 助詞の分布

書き手の特徴を抽出するための単語に関する計量方法としては，単語の長さの分布による統計分析方法がよく用いられている．しかし，助詞の場合は，単語の長さが極端に短

く，頻繁に使われている助詞はほとんどが1～2文字である．したがって，助詞の場合は，長さによる計量方法は書き手の特徴が現れにくい（金1994 a）．都合よく頻繁に使われている助詞は約20種類前後である．そこで，本研究では出現頻度が高い助詞23種類及びその他の助詞を1つにまとめた合計24項目に分け計量分析することにした．ただし，本研究では助詞について詳細な分類は行わなかった．表3に24項目に分けたそれぞれの使用頻度を示した．

表3．各文章における助詞・項目毎の出現頻度と出現率(1)

ID	か	が	て	で	と	に	の	は	ば	へ	も	や
I 1	19	125	218	82	160	168	321	199	12	26	63	4
I 2	24	121	217	73	144	148	288	219	11	27	75	0
I 3	19	147	209	50	151	176	320	183	14	15	55	5
I 4	23	133	204	62	129	154	329	205	6	20	44	2
I 5	9	101	156	39	113	152	257	189	13	33	44	5
I 6	21	122	203	48	84	153	239	157	2	21	35	5
I 7	39	140	174	37	100	110	263	147	5	38	26	0
I 8	25	138	241	52	123	149	313	192	1	27	45	4
M 1	12	150	186	48	86	193	393	248	9	25	28	9
M 2	8	139	184	56	86	135	274	170	6	23	40	2
M 3	15	146	171	38	120	178	341	227	12	15	54	15
M 4	6	101	166	47	55	152	275	176	2	29	31	8
M 5	12	177	280	67	76	262	476	252	10	33	26	13
M 6	6	149	237	63	100	200	410	236	8	34	49	7
M 7	10	198	235	90	113	254	462	293	2	36	50	16
M 8	18	173	267	70	107	275	446	280	10	31	59	13
M 9	14	195	224	75	129	269	414	271	8	23	56	8
M10	16	182	205	71	106	236	428	268	8	28	48	19
M11	14	175	226	59	100	245	422	247	7	27	56	9
N 1	23	84	129	35	80	167	230	127	15	3	41	2
N 2	11	77	125	22	87	167	237	134	13	5	30	0
N 3	13	121	168	34	110	172	301	165	15	5	48	9
N 4	15	132	134	32	133	174	330	183	5	9	55	4
N 5	23	121	144	26	138	182	272	142	20	5	34	2
N 6	13	115	191	24	132	228	333	175	16	27	48	5
N 7	19	130	184	23	135	229	335	192	13	9	54	9
N 8	19	124	160	30	118	210	340	204	13	14	50	12
N 9	18	139	181	40	122	216	330	174	13	9	62	8

表3のつづき

ID	を	から	だけ	ても	でも	とも	ので	ほど	まで	ながら	ばかり	その他
I 1	8	36	14	10	11	4	15	11	20	1	4	32
I 2	193	36	15	7	11	5	2	12	10	6	4	38
I 3	177	37	9	3	6	3	4	6	13	7	1	49
I 4	158	35	5	4	8	2	4	6	9	10	1	32
I 5	129	45	3	8	9	0	10	1	8	2	3	37
I 6	152	27	5	4	5	2	10	0	2	7	6	48
I 7	157	41	9	6	10	2	3	3	8	4	2	38
I 8	159	48	8	1	5	0	3	1	14	5	1	30
M 1	220	30	3	3	1	1	22	3	2	9	3	37
M 2	180	25	7	3	0	8	27	6	12	6	5	52
M 3	180	31	4	6	8	2	12	3	2	6	1	61
M 4	149	29	0	5	2	1	9	2	4	1	0	23
M 5	257	48	6	3	5	1	32	0	13	5	4	33
M 6	263	52	4	6	3	0	23	3	14	9	2	42
M 7	299	41	8	1	5	2	36	2	10	7	1	49
M 8	294	41	6	7	3	5	42	3	11	12	6	35
M 9	255	44	9	10	8	4	30	2	13	8	1	44
M10	234	36	12	2	4	2	37	1	4	4	3	13
M11	310	32	7	1	3	2	32	1	6	8	2	38
N 1	130	28	1	3	12	2	0	0	4	6	6	31
N 2	176	20	1	6	5	3	1	7	4	5	1	34
N 3	165	38	7	10	8	5	3	12	12	1	5	59
N 4	173	24	7	11	5	3	3	5	7	5	2	44
N 5	175	23	3	5	5	1	2	3	5	6	6	43
N 6	213	43	2	10	6	3	0	5	15	3	4	50
N 7	178	21	7	15	3	7	1	6	12	3	4	48
N 8	227	22	15	10	7	5	3	4	12	9	5	51
N 9	207	32	3	5	6	3	0	3	13	4	6	40

いま，文章 i における助詞 j の使用頻度を x_{ij} と表すと，全体で I 編の文章における，合計の項目数が J としたときの助詞の使用頻度と相対使用率のマトリックスはそれぞれ

$$X_{I\times J}=[x_{ij}]$$

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

$$p_{ij}=\frac{x_{ij}}{\sum_{v=1}^{n}x_{iv}},\quad \sum_{j=1}^{n}p_{ij}=1$$

と表示できる．

本文では相対使用率に関するデータを分布と呼ぶことにする．

4．2 分布間の距離

以下では，同じ書き手の，二つの分布の間における距離を群内距離，異なる書き手の二つの分布の間における距離を群間距離と呼ぶ．もし，助詞の分布に書き手の特徴が現れているとすると，群内距離の平均値が群間距離の平均値より小さいはずである．群内距離の平均値が群間距離の平均値より小さければ小さいほど書き手の特徴が明確である．

本研究では，文章 i の助詞の分布と文章 l の助詞の分布の間の距離 d_{il} は次の式を用いて求める（*McLachlan* 1992）．

$$d_{il}=\frac{1}{2}\sum_{j=1}^{n}\left(p_{ij}\log\frac{2p_{ij}}{p_{ij}+p_{lj}}+p_{lj}\log\frac{2p_{lj}}{p_{ij}+p_{lj}}\right)$$

ただし

$$p_{ij}=0\text{ なら }p_{ij}\log\frac{2p_{ij}}{p_{ij}+p_{lj}}=0$$

$$p_{lj}=0\text{ なら }p_{lj}\log\frac{2p_{lj}}{p_{ij}+p_{lj}}=0$$

とする．以下では上式により求めた距離を *K-L-S* 距離と呼ぶ．

K-L-S 距離により求めた分布の間の距離マトリックスを

$$D_{I\times I}=\begin{bmatrix}0 & d_{12} & \cdots & d_{1I}\\ d_{21} & 0 & \cdots & d_{2I}\\ \vdots & \vdots & \ddots & \vdots\\ d_{I1} & d_{I2} & \cdots & 0\end{bmatrix}$$

で表記する．

いま，書き手 k と h のそれぞれ k_n，h_m 編の文章があったときに，それぞれ書き手に関し，群内，群間での任意の二つの文章の，すべての組合せに対して，助詞の分布の間の距離の平均をそれぞれ

$$\overline{d(k)}=\frac{2\sum_{k_i=k_1}^{k_n-1}\sum_{k_j=k_i+1}^{k_n}d_{k_ik_j}}{(k_n-1)k_n}\times 100$$

$$\overline{d(h)}=\frac{2\sum_{h_i=h_1}^{h_m-1}\sum_{h_j=h_i+1}^{h_m}d_{h_ih_j}}{(h_m-1)h_m}\times 100$$

$$\overline{d(k,h)}=\frac{\sum_{k_i=k_1}^{k_n}\sum_{h_j=h_1}^{h_m}d_{k_ih_j}}{k_nh_m}\times 100$$

で求めた．

表 3 のデータから，3 人の 28 の文章における助詞の分布を用いて求めた群内距離，群間距離の平均値を表 4 に示した．

表 4．助詞の分布間の距離の平均値

書き手名	群内	最小の群間	群間		
			井上	三島	中島
井上	0.876	1.386	0.876	1.401	1.386
三島	0.685	1.401	1.401	0.685	1.681
中島	0.655	1.386	1.386	1.681	0.655

表 4 から，3 人の群内距離の平均値が最小の群間距離の平均値より小さいことがわかる．この結果から，助詞の分布には書き手の特徴が現れると判断する．

ちなみに，群内距離の平均値と最小の群間距離の平均値を用いて分散分析を行った．その結果，分散比 F 値は 191.60 で，p 値は 0.0052

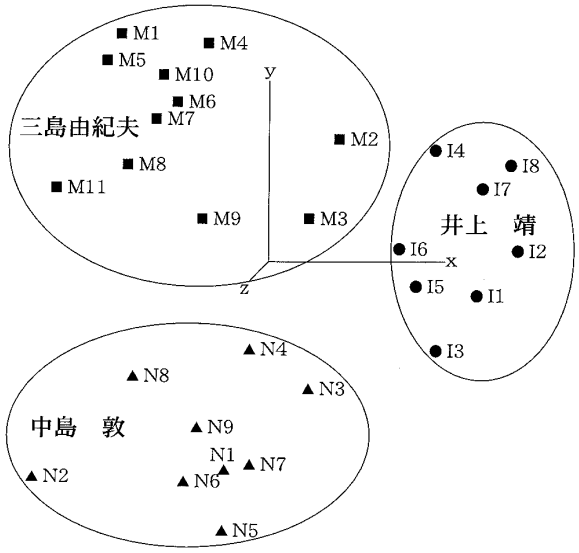
である．この値から群内距離の平均値と最小の群間距離の平均値との間には明らかな差があると判断する．これは助詞の分布に書き手の特徴が比較的明確に現れていることを意味する．

4．3　文章の分類

前節の分析で，助詞の分布に書き手の特徴が現れていることが分かった．文章の書き手を識別するためには，書き手の特徴情報を用いて文章を分類した場合，文章が書き手ごとに分類されることが望まれる．本節では，助詞の分布を用いて文章を分類する場合，文章が書き手ごとに分類されるか否かという視点から書き手の特徴について分析を行う．

分類分析には多くの方法があるが，本研究では広く知られている主成分分析方法を用いる．主成分分析は $P_{I\times J}$ の分散共分散の行列を用いて行う．第1主成分，第2主成分，第3主成分の寄与率はそれぞれ31.15%，29.29%，10.21%で，第3主成分までの累積寄与率は70.65%である．図1に見やすい角度に回転した第1～第3主成分得点の三次元散布図を示した．主成分得点の3次元散布図では，文章が書き手別に分類されることがわかる．

図1．助詞の分布に基づいた文章の3次元散布図



4．4　書き手の識別

前節では学習データなしのアプローチで，文章の分類を行う場合，文章が書き手ごとに分類されるか否かについて分析を行なった．本節では，学習データがあるという仮定のもとで，文章の書き手がどの程度の確率で正しく識別されるかについて分析を行う．

判別分析にも多くの方法が提案されている．本研究では，重判別分析（正準判別分析）の一種である，距離による判別分析法を用いることにする．*K-L-S* 距離を用いて，表1に示したすべてのデータのなかから一つの個体（文章）を除いて学習を行ない，除かれた文章がどの書き手に属するかについて判別を行なった．その結果を表5に示す．表5から分かるようにすべての文章が正しく識別され，その識別率は100%である．

主成分分析の3次元散布図と対応するため，第1から第3までの主成分得点を用いてマハラノビス距離による判別分析も試みた．その結果 *K-L-S* 距離を用いた場合と同一な結果が得られた．

表5．助詞の分布を用いた判別結果

書き手名	井　上	三　島	中　島
井　上	8	0	0
三　島	0	11	0
中　島	0	0	9

4．5　書き手の識別ルールの抽出

助詞分布における書き手ごとの特徴を見つけ出すためには，

(1) 分布の変数（項目）ごとの t, F, χ^2 等の統計量

(2) 主成分分析，因子分析，対応分析の個体のスコアと変数のスコアの対応

を用いて分析を行うことが考えられる．しかし，このような方法は分布の中のどの変数（項目）に書き手の特徴がより明確に現れるかについて考察することは可能であるが，書き手を識別するルールを抽出するまでには

至っていない．そこで本章では近年注目を集めつつあるデータマイニング法を用いて助詞分布における書き手を識別するルールの抽出を試みることにする．データマイニング法も様々なアプローチで研究が進められているが，ここではラフ集合理論に基づいたデータマイニング法を用いる．

ラフ集合（*rough sets*）の概念はその歴史が浅いため，広く知られていない．そこで，まずラフ集合について簡単に記しておく．ラフ集合はポーランドの計算機科学者 *Pawlak* が 1982 年提案した（*Pawlak* 1982）．ラフ集合の概念は集合上の「類別」と「近似」である．ラフ集合の定義を与えるため，全体集合 $U=\{S_1, S_2, \cdots, S_n\}$ 上に集合 X があるとする．いま S_i を用いて X を近似することを考える．S_i を用いた X の近似は

(1) X に完全に含まれている S_i を用いて X を近似

(2) X を含む S_i を用いて X を近似

の二つの方法がある．方法(1)による X への近似は下近似，方法(2)による X への近似は上近似と呼び，それぞれ $\underline{A}X, \overline{A}X$ と記する．X の下近似値と上近似値が一致しない場合，集合 X をラフ集合という．ラフ集合では，属性集合における下記のような下近似と上近似の関係式

$$\alpha_A(X)=\frac{Card(\underline{A}X)}{Card(\overline{A}X)}$$

を用いて近似の度合を示す．式のなかの $Card(\underline{A}X)$，$Card(\overline{A}X)$ はそれぞれ下近似，上近似集合のなかに含まれた要素の数である．明らかに $\alpha(X)$ は $0 \leqq \alpha(X) \leqq 1$ であり，$\alpha(X)$ が大きいほど近似の精度がよい．

このような概念と *Ziarko*(1993) のヒントのもとで，分類を行うときの判別精度の度合を下記のように定義する．

$$E(Bi,Ci)=\begin{cases} 1-\frac{1}{2}\left(\frac{Card(Bi,\cap Ci)}{Card(Bi)}+\frac{Card(Bi,\cap Ci)}{Card(Ci)}\right) \\ \quad if\ Card(Bi)>0\ or\ Card(Ci)>0 \\ 0 \quad if\ Card(Bi)=0\ or\ Card(Ci)=0 \end{cases}$$

$Card(B_i), Card(C_i)$ は，属性（変数）i におけるそれぞれ B, C に含まれる要素の数であるが，本稿では，用いた著者 B, C の文章の数に等しい．$Card(B_i \cap C_i)$ は，属性（変数）i における B の要素が C の範囲のなかに含まれる，あるいは C の要素が B の範囲のなかに含まれる数である．$E(B_i, C_i)$ の値を閾値という．閾値が高いほど分類の度合がよいことになる．識別・判別のルールは閾値が高い上位ランクの変数（項目）の組合せにより構成することにする．

識別ルールの生成を行うためには，まず閾値を決める必要がある．閾値は試行錯誤でデータと対話しながら決める必要がある．

助詞分布の各変数（項目）の p_{ij} は小数点以下の数値が多いため $P_{I\times J}$ に 100 を乗じ，小数点以下 2 桁まで用いた．28 文章における助詞分布データでは閾値を 0.85 とすると，井上と三島，井上と中島が識別できる．助詞として「か」，「て」，「と」，「に」が候補として選択される．これらを用いて，プログラムにより生成された，最も少ない変数（助詞），かつ最も高い識別率で井上と他の作家を識別する知識は下記のルール 1，2 である．

ルール 1 *if* $(11.42<x_3)$ *and* $(6.19<x_5)$ *and* $(x_6<11.27)$ *then* 井上

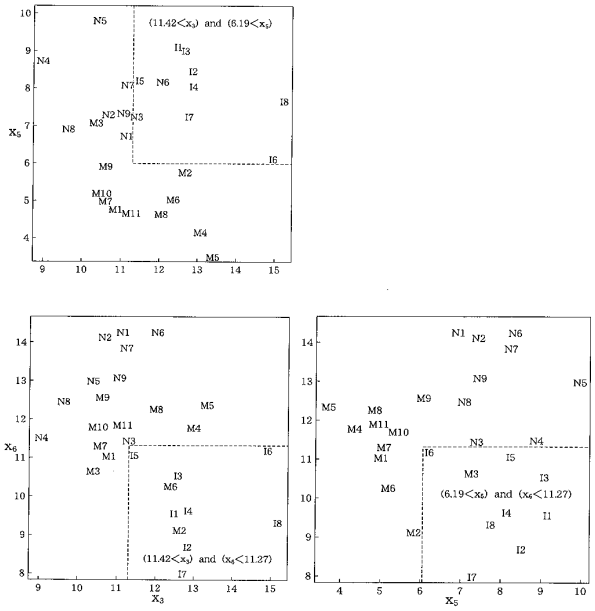
ルール 2 *if*$(0.66<x_1<2.86)$*and*$(11.42<x_3<15.21)$*and*$(x_6<11.27)$*then* 井上

ルールの中の x_1, x_3, x_5, x_6 はそれぞれ「か」，「て」，「と」，「に」の使用率（助詞の中に占めるパーセンテージ）である．上記の二つのルールによる識別率はいずれも 100% である．またここで抽出されたルールはいずれも

井上と三島・中島を同時に識別可能なルールである．ちなみに，井上と中島だけを識別するなら$x_6<11.27$だけで十分である．

図2に井上を識別するルール1に用いた3つの変数による対散布図を示す．点線で囲まれた部分は2つの変数の条件を満たす井上の領域である．

図2．3つの助詞を用いた対散布図



同様な方法で求めた三島，中島を識別するルールの一部は下記の通りである．

ルール3 *if* $(x_1<0.91)$ *and* $(3.63<x_5<7.28)$ *and* $(0.71<x_{19})$ *then* 三島

ルール4 *if* $(11.57<x_6<14.41)$ *and* $(x^{19}<0.20)$ *then* 中島

ルールの中のx_{19}は「ので」の使用率である．このようなルールは多く存在するがここに抽出されたルールは最大の識別率でかつ最も少ない変数（項目）を用いたルールである．このようなルールは書き手ごとの特徴を捉まえるためには多いに役立つであろう．

8．終わりに

本研究では，文章の中で出現頻度が最も高い助詞に着眼し，三人の文章を用いて，

(1) 助詞の分布には書き手の特徴が現れるか否か

(2) 助詞分布を用いて，著者を識別・判別する際の識別率はどの程度であるか

(3) 助詞の分布から書き手を識別する簡潔なルールを抽出することが可能であるか否かについて計量分析を試みた．

その結果，助詞の分布には書き手の特徴が明確に現れること，および助詞の分布は，書き手の文章を分類する，あるいは文章の書き手の識別・判別を行うための有力な書き手の特徴情報であることが明らかになった．また助詞の分布には書き手の特徴が明確に現れるため，データマイニング法により書き手を識別・判別するルールを抽出し，書き手ごとの特徴を明らかにすることが可能であることがわかった．識別ルールの識別・判別率は判別分析より劣ることは自明であるが，そのルールは書き手の特徴を掴むための有益な情報となる．

助詞は比較的使用率が高いため，短い文章の書き手の識別・判別においては，助詞の分布は，動詞の長さの分布（金 1994 *a*，1995，1996 *a*），読点の打ち方に関する情報（*Jin and Murakami* 1993，金・他 1993 *a*，金 1994 *a*，1994 *b*）より有効である．今後の課題としては，助詞の分布に現れる書き手の特徴に関して言語学，心理学観点からの裏付けがあげられる．

謝辞

本研究で用いたタグ付文章のデータは，文部科学省統計数理研究所村上征勝教授の研究室で作成したものである．本研究は札幌学院大学社会情報学部理系プロジェクトの研究助成金を受けた．また本稿の査読者から貴重なご意見を頂戴した．記して感謝します．

参考文献

Holmes,D.I.(1994).Authorship Attribution, Computers and the Humanities, 28, 87-106.

Jin,M and Murakami, M (1993). Authors' Characteristic Writing Styles as Seen Through Their Use of Commas, Behaviormetrika, Vol. 20, 63-76.

金明哲，樺島忠夫，村上征勝（1993 a）．読点と書き手の個性，計量国語学，Vol. 18, No. 8, 382-391.

金明哲，樺島忠夫，村上征勝（1993 b）．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, No. 3, 133-145.

金明哲（1994 a）．自然言語におけるパターンに関する計量的研究，総合研究大学院大学，学位論文.

金明哲（1994 b）．読点の打ち方と文章の分類，計量国語学，Vol. 19, No. 17, 317-383.

金明哲（1995）．動詞の長さの分布に基づいた文章の分類と和語および合成語の比率，自然言語処理，Vol. 2, No. 1, 57-75.

金明哲（1996 a）．動詞の長さの分布と文章の書き手，社会情報，札幌学院大学社会情報学部紀要，Vol. 5, No. 2, 13-22.

金明哲（1996 b）．小説文における文節の係り受け距離の統計的特徴，計量国語学，Vol. 20, No. 4, 168-179.

金明哲（1996 c）．文節の係り受け距離の統計的分析，社会情報，Vol. 5, No. 2, 1-12.

金明哲（1996 d）．助詞分布に基づいた文章の書き手の認識，人文科学における数量的分析論文集（文部省科学研究費・重点領域研究），49-54．行動計量学会第 24 回大会論文抄録集，144-147.

McLachlan, G. J. (1992). Discriminant Analysis and Statistical Pattern Recognition, JOHN WILEY, SONS, INC.

韮沢正（1965）．由良物語の書き手の統計的判別，計量国語学，No. 33, 21-28.

村上征勝（1994）．計量的文体研究の威力と成果，言語，Vol. 23, No. 2, 30-37.

村上征勝・金明哲（1998）．「人文科学とコンピュータ」講座第 5 巻「数量的分析編」，尚学出版.

Pawlak,Z. (1982):RoughSets. International Journal of Computer and Information Sciences, 11, 341-356.

Pawlak, Z. (1984): Rough Classification, Int. J. Of Man-Machin Studies, 20, 469-485.

佐々木和枝（1976）．文の長さの分布型，計量国語学，No. 78, 13-22.

安本美典（1957 a）．文の長さの分布型，計量国語学，No. 1, 20-30.

安本美典（1957 b）．文体統計各種分布型，計量国語学，No. 2, 20-24.

安本美典（1994）．文体を決める三つの因子，言語，Vol. 23, No. 2, 22-29.

Ziarko, W. (1993). Variable Precision rough Sets Model. Journal of Computer and System Science, Vol. 46, No. 1, 29-59.

2002 年 1 月 28 日受付
2002 年 2 月 15 日受理