Vol.2017-CH-113 No.1
2017/2/4

ランダムフォレストを用いた文芸テキスト分類と指標の評価 — Alice Bradley Sheldon と Ernest Hemingway —

木村美紀[†1]a

概要：本研究では，ランダムフォレストを用いた文芸テキストの分類を行う．分析対象は Alice Bradley Sheldon 全 72 作品と Ernest Hemingway の短編小説全 69 作品である．高頻度語彙上位 50 語という指標を用いて Alice Bradley Sheldon と Ernest Hemingway のランダムフォレストによるテキスト分類を行った結果, 92.20%という高い分類正確率を得ることができた．本研究では，先行研究の変数を変更して分析を行い先行研究の結果が追試可能か検討を行う．Hirst and Feiguina (2007) , Hou and Jiang (2016) 使用されているような品詞の分布という統語的な指標を追加し分析を行った．

キーワード：ランダムフォレスト，Alice Bradley Sheldon，Ernest Hemingway，計量文体論

Quantitative Authorship Attribution of Two Authors and the Evaluation of Variables — Alice Bradley Sheldon and Ernest Hemingway —

MIKI KIMURA[†1]

Abstract: This is a case study on the quantitative authorship attribution of the works of two writers, Alice Bradley Sheldon and Ernest Hemingway. As variables, based on Hirst and Feiguina (2007) and Hou and Jiang (2016), the distribution of the parts-of-speech is selected. In investigating the dissimilarities between the works by Alice Bradley Sheldon and those by Ernest Hemingway, this study employs random forests.

Keywords: Random Forests, Alice Bradley Sheldon, Ernest Hemingway, Quantitative Stylistics

1. はじめに

本研究では，コーパス言語学の一分野である計量文体論の手法を用いて，文芸批評上比較されることの多い Alice Bradley Sheldon 作品群と Ernest Hemingway 作品群のテキスト分類を行う． Alice Bradley Sheldon (1915-1987: 米国) は James Tiptree, Jr. と Raccoona Sheldon という 2 名義を使用して正体不明・性別不明の作家として著作活動を行っていた作家である．主に短編の SF 小説を執筆していた作家で，特に文芸批評において Alice Bradley Sheldon の男性名義である James Tiptree, Jr. 名義作品群は Ernest Hemingway 作品群との比較されることが多い．

本研究ではランダムフォレストを用いて Alice Bradley Sheldon 作品群と Ernest Hemingway 作品群の 2 群の分類を試みる． 本稿では，ランダムフォレストに基づく多次元尺度法によるテキスト分類と同時に Gini 係数に基づく分類に有効であった変数の提示を行う． さらに，部分従属プロットによって分類に有効であった変数の分布を提示した．

変数としては, Hirst and Feiguina (2007) や Hou and Jiang (2016) で使用され，テキスト分類においてその有効性が示されている品詞分布を採用した．

2. 先行研究

本章では, Alice Bradley Sheldon や Sheldon の男性名義である James Tiptree, Jr. 作品群に関する文芸批評上の文体評価と，この著者に対して計量文体論の手法を用いて分析を行った論文執筆者自身の先行研究を提示する．

2.1 文芸批評における文体比較

Silverberg [1] では， “It has been suggested that Tiptree is female, a theory that I find absurd, for there is to me something ineluctably masculine about Tiptree’s writing. I don’t think the novels of Jane Austen could have been written by a man nor the stories of Ernest Hemingway by a woman, and in the same way I believe the author of the James Tiptree stories is male.” のように Ernest

†1 明治大学大学院文学研究科博士後期課程
Graduate School of Arts and Letters, Meiji University

a) mk_ling@meiji.ac.jp

Hemingway に言及しながら正体不明・性別不明の作家として著作活動を行っていた James Tiptree, Jr. に対する性別推定が行われている.

また, 小谷 [2] では,「ジェイムズ・ティプトリー・ジュニアなる作家は, (中略), その華麗な文体, ヘミングウェイを思わせるマッチョな作風で一躍 SF 界を魅了した. 時代はフェミニズム SF 華やかなりし頃, そのなかでこの著者不明 = 正体不明の作品は, その作風から, 手堅い稀有の才能を持つ男性新人作家の書いたものと判断されていた」というように, Ernest Hemingway との比較を行っている.

2.2 計量文体論の手法を用いた文体比較

Alice Bradley Sheldon 作品群と Ernest Hemingway 作品群に対し計量文体論手法を用いて分類を行った先行研究としては, 論文執筆者自身が実施した Kimura [3] が挙げられる.

Kimura [3] では, その実績と有効性から Burrows [4], Burrows [5], Burrows and Hassal [6] などで採用されている「高頻度語彙上位 50 語」と「品詞分布」を指標として採用した. 統計手法としては, 教師なしの分類手法であるマルチスケールブートストラップ法に基づくクラスター分析と主成分分析 (PCA) , 教師ありの分類手法である判別分析とサポートベクターマシン (SVM) を用いた. これらの統計手法を用いて, Alice Bradley Sheldon 男女 2 名義 (James Tiptree, Jr. と Raccoona Sheldon) 作品群と Ernest Hemingway 作品群の 3 カテゴリーでの分類を検証した.

高頻度語彙上位 50 語を用いて検証を行ったところ, クラスター分析と主成分分析では Alice Bradley Sheldon 作品群における名義での文体差は確認できなかった. 一方 Alice Bradley Sheldon 作品群と Ernest Hemingway 作品群は完全に分離が可能であった. 図 1 には主成分分析の出力を示す.

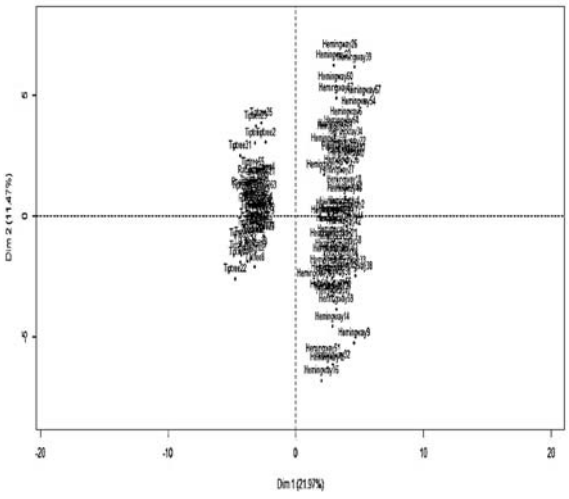

図 1 主成分分析二次元散布図 (相関行列)

図 1 の散布図において x 軸の負の方向に布置しているのが Alice Bradley Sheldon 作品群であり, x 軸の正の方向に布置いるのが Ernest Hemingway 作品群である.

また, 教師ありの分類手法である判別分析と SVM の分類正確率はそれぞれ 93.66%と 96.49%であった. これらの正確率はサンプルサイズに基づいた分類正確率の基準を有意に上回っている.

次に, 品詞分布に基づいた分類結果を示す. 品詞分布を指標として採用した場合では, 教師なしの分類手法 2 種では判別が不可能だった. 教師ありの分類手法である判別分析と SVM の分類正確率はそれぞれ 88.03%と 95.77%であった. これらの正確率はサンプルサイズに基づいた分類正確率の基準を有意に上回っている. このように, Alice Bradley Sheldon 作品群と Ernest Hemingway 作品群は分類可能であると結論付けられる.

3. 品詞分布に基づく計量文体分析

3.1 分析に用いたデータ

本研究では以下の 12 冊の紙媒体の書籍を電子化し Alice Bradley Sheldon (延べ 865,802 語) コーパスを構築した. 表 1 に Alice Bradley Sheldon コーパス構築に使用した底本を示す.

表 1. Alice Bradley Sheldon 底本

	作品名	出版年
1	*Brightness Falls from the Air*	1993
2	*Byte Beautiful: Eight Science Fiction Stories*	1985
3	*Crown of Stars*	1988
4	*Her Smoke Rose Up Forever*	1990
5	*Meet Me at Infinity*	2002
6	*Out of the Everywhere and Other Extraordinary Visions*	1981
7	*Star Songs of an Old Primate*	1978
8	*Tales of the Quintana Roo*	1986
9	*Ten Thousand Light-Years from Home*	1973
10	*The Starry Rift*	1986
11	*Up the Walls of the World*	1978
12	*Warm Worlds and Otherwise*	1975

本研究では, 表 1 で示した底本から構築したコーパスに含まれている Alice Bradley Sheldon 全 72 作品を分析対象とした. また, Ernest Hemingway 作品群に関しては, *The Complete Short Stories of Ernest Hemingway* を底本として Ernest Hemingway 短編作品全 69 作品を含むコーパス (延べ 271,475 語) を構築し, 分析対象とした.

本研究においては, Hirst and Feiguina [7] や Hou and Jiang [8] に基づきテキスト中の品詞分布を指標として採用した. 本研究では Gotagger というソフトウェアを使用し, 品詞のアノテーションを行った. 図 2 に *Brightness Falls from the*

Air に品詞タグのアノテーションを行った Gotagger の出力例を提示する．また, Gotagger の品詞タグセットを表 2 に示す．

表 2　Gotagger 品詞タグセット

品詞タグ略号	品詞
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition/subord. conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund/present participle
VBN	Verb, past participle
VBP	Verb, non-3rd ps. sing.
VBZ	Verb, 3rd ps. sing. Present
WDT	wh-determiner
WP	wh-pronoun
WP$	Possessive wh-pronoun
WRB	wh-adverb

```
In_IN the_DT driver_NN 's_POS seat_NN beside_IN her_PRP$ ,_, Kipruget_NNP Korso_
NNP known_VBN to_TO all_DT as_IN Kip_NNP squints_NNS up_IN at_IN the_DT descendi
ng_VBG fires_NNS ._. He_PRP is_VBZ Deputy_NNP Administrator_NNP and_CC Dameii_NN
P Guardian_NNP -_: Liaison_NNP ,_, as_RB well_RB as_IN Corys_NNP mate_NN ._.↓
Cory_NNP 's_POS brown_JJ eyes_NNS slide_VBP sideways_RB to_TO him_PRP ,_, and_CC
 she_PRP smiles_VBZ ._. Kip_NNP is_VBZ the_DT handsomest_JJS man_NN she_PRP 's_V
BZ ever_RB seen_VBN ._. a_DT fact_NN of_IN which_WDT he_PRP seems_VBZ quite_RB u
naware_JJ ._.↓
```

図 2　品詞タグ付与例

Gotagger の品詞情報に基づき，品詞分布，品詞の bigram，品詞の trigram という 3 種の変数を抽出した．また，変数の dispersion を検討し，「10 種以上のテキストに出現していない変数」の削除行った．

3.2 分析

本研究では, Breiman [9] でその手法が提唱され，金・村上 [10] において様々なジャンルのテキスト分類において，SVM 等の既存の分類法に比べて分類感度が高いと結論付けられているランダムフォレストを分類手法として採用した．

3.2.1 品詞分布

品詞分布を用いて行ったテキスト分類の結果を表 2 に示す．表 3 から, Alice Bradley Sheldon72 作品中 66 作品が正しく分類されているということが分かる．また，Ernest Hemingway69 作品中 66 作品が正しく分類されているということが判明した．今回分析に使用したデータセットに対する分類正確率は 93.62%だった．これは，サンプルサイズを基準とした分類正確率を有意に上回っている．

表 3 ランダムフォレスト出力

	Alice Sheldon	Ernest Hemingway
Alice Sheldon	**66**	6
Ernest Hemingway	3	**66**

図 3 には Gini 係数に基づく変数の特徴度を提示した．図 3 から，VBG (Verb, gerund/present participle), NNS (Noun, plural), JJ (Adjective), DT (Determiner) などが特徴的な変数として挙げられる．

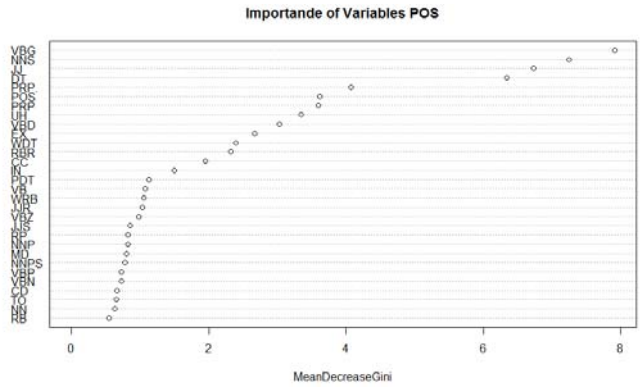

図 3　Gini 係数に基づく変数の特徴度

Cutler et al. [11] に基づき部分従属プロットの出力を検証する．図 4 では縦軸の値が大きく，右肩上がりの折れ線グラフであると，その変数が第 1 群 (Ernest Hemingway 作品群) において特徴的であると結論付けられる．反対に縦軸の値が小さく右肩下がりになっている折れ線グラフであると，その変数が第 2 群 (Alice Bradley Sheldon 作品群) において特徴的であると結論付けられる．

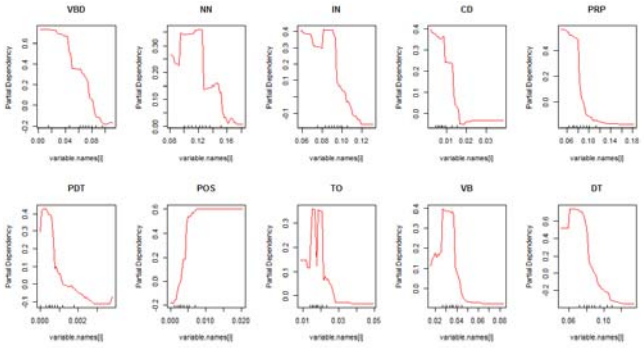

図 4　部分従属プロット

具体的には，変数 VBD は縦軸の値が小さいため第 2 群である Alice Bradley Sheldon 作品群に特徴的な変数であると結論付けられる．一方，変数 POS は縦軸の値が大きいため第 1 群である Ernest Hemingway 作品群に特徴的な変数であると結論付けられる．このようにして各変数の特徴度を検証していくと，NN, IN, CD, PRP という変数が第 2 群である Alice Bradley Sheldon 作品群において特徴的であると結論付けられる．このようにして，図 3 においてテキストの判別に寄与している変数が Alice Bradley Sheldon 作品群と Ernest Hemingway 作品群という 2 群においてどちらに特徴的な変数であるのかということが同定可能になった．

3.2.2　品詞 bigram

品詞の bigram を指標として採用して行ったテキスト分類の結果を表 4 に示す．表 4 から，Alice Bradley Sheldon72 作品中 66 作品が正しく分類されているということが分かる．また，Ernest Hemingway69 作品中 67 作品が正しく分類されているということが判明した．今回分析に使用したデータセットに対する分類正確率は 94.33%だった．これは，サンプルサイズを基準とした分類正確率を有意に上回っている．

表 4　bigram ランダムフォレスト出力

	Alice Sheldon	Ernest Hemingway
Alice Sheldon	**66**	6
Ernest Hemingway	2	**67**

図 5 には Gini 係数に基づく変数の特徴度を提示した．図 5 から，PRP JJ (Personal pronoun と Adjective), JJ NN (Adjective と Noun, singular or mass), JJ JJ (Adjective と Adjective), IN JJ (Preposition/subord. Conjunction と Adjective), JJ NNS (Adjective と Noun, plural) などが特徴的な変数として挙げられる．

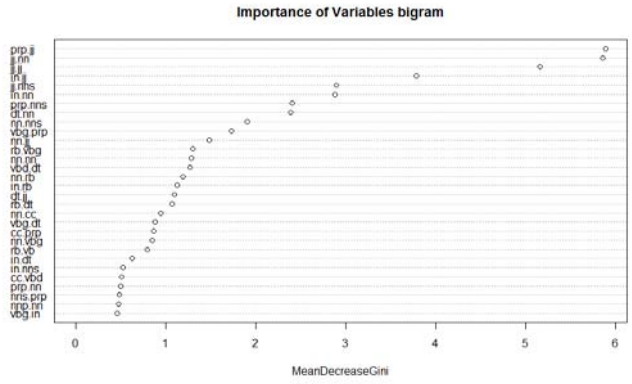

図 5　Gini 係数に基づく変数の特徴度

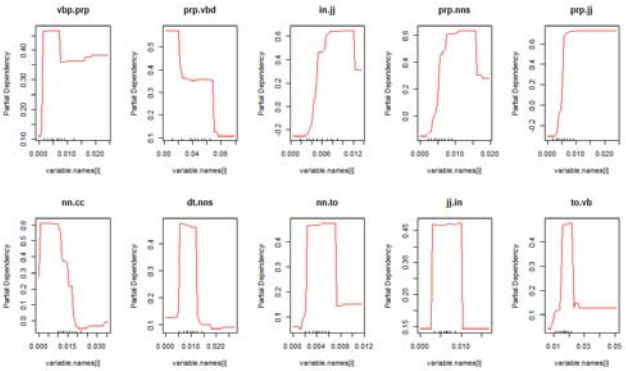

図 6　部分従属プロット

図 6 によって，部分従属プロットを用いて特徴的な変数の分布を検証していく．具体的には，変数 VBP PRP は縦軸の値が小さいため第 2 群である Alice Bradley Sheldon 作品群に特徴的な変数であると結論付けられる．一方，変数 PRP VBD は縦軸の値が大きいため第 1 群である Ernest Hemingway 作品群に特徴的な変数であると結論付けられる．このようにして各変数の特徴度を検証していくと，NN CC, DT NNS という変数が第 2 群である Alice Bradley Sheldon 作品群において特徴的であると結論付けられる．このようにして，部分従属プロットを用いることによって図 5 においてテキストの判別に寄与している変数が Alice Bradley Sheldon 作品群と Ernest Hemingway 作品群という 2 群においてどちらに特徴的な変数であるのかということが同定可能になった．

3.2.3　品詞 trigram

表 5　bigram ランダムフォレスト出力

	Alice Sheldon	Ernest Hemingway
Alice Sheldon	**65**	7
Ernest Hemingway	7	**62**

品詞の trigram を指標として採用して行ったテキスト分類の結果を表 5 に示す. 表 4 から, Alice Bradley Sheldon72 作品中 65 作品が正しく分類されているということが分かる. また, Ernest Hemingway69 作品中 62 作品が正しく分類されているということが判明した. 今回分析に使用したデータセットに対する分類正確率は 90.07%だった. これは, サンプルサイズを基準とした分類正確率を有意に上回っている.

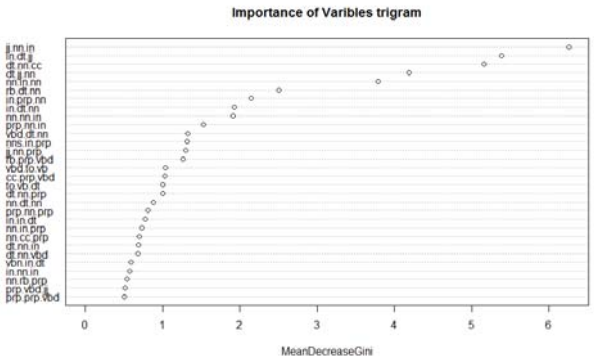

図 7　Gini 係数に基づく変数の特徴度

図 7 には Gini 係数に基づく変数の特徴度を提示した. 図 7 から, JJ NN IN, IN DT JJ, DT JJ NN, NN IN NN, RB DT NN などが特徴的な変数として挙げられる.

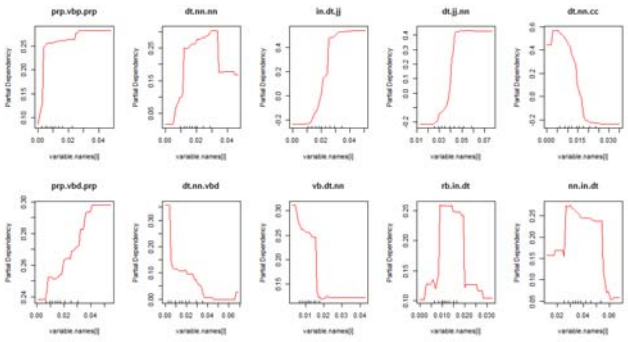

図 8　部分従属プロット

図 8 の部分従属プロットを用いて特徴的な変数の分布を検証していく.変数 DT NN CC, DT NN VBD, VB DT NN は縦軸の値が小さいため第 2 群である Alice Bradley Sheldon 作品群に特徴的な変数であると結論付けられる. 一方, 変数 PRP VBR PRP, DT NN NN, IN DT JJ, DT JJ NN は縦軸の値が大きいため第 1 群である Ernest Hemingway 作品群に特徴的な変数であると結論付けられる.このようにして, 部分従属プロットを用いることによって図 7 においてテキストの判別に寄与している変数が Alice Bradley Sheldon 作品群と Ernest Hemingway 作品群という 2 群においてどちらに特徴的な変数であるのかということが同定可能になった.

4. 結論

前章で行った分析から, 品詞分布, 品詞の bigram, 品詞の trigram を指標とした試行での分類正確率はそれぞれ 93.62%, 94.33%, 90.07%だった. これらの分類正確率は, サンプルサイズを考慮に入れた正確率の基準を有意に上回っている. 今回のデータセットに関しては, 品詞の bigram という指標が分類に有効であったということが判明した.

さらに, Gini 係数に基づく変数の特徴度の提示や部分従属プロットを用いて Alice Bradley Sheldon 作品群と Ernest Hemingway 作品群の分離に有効であった変数の特定とその分布を検証した.

本研究での試行から, 小谷 [2] で指摘されている, 「ジェイムズ・ティプトリー・ジュニアなる作家は,(中略), その華麗な文体, ヘミングウェイを思わせるマッチョな作風で一躍 SF 界を魅了した. 時代はフェミニズム SF 華やかなりし頃, そのなかでこの著者不明 = 正体不明の作品は, その作風から, 手堅い稀有の才能を持つ男性新人作家の書いたものと判断されていた」 という Alice Bradley Sheldon 作品群と Ernest Hemingway 作品群の文体の類似性というのは品詞の分布から来ているわけではないかもしれない.

謝辞　論文作成に際し有益なコメントをくださったミシシッピ大学所属の生田敏一准教授に謹んで感謝の意を表す.

参考文献

[1] Silverberg, R. Who Is Tiptree, What Is He? *Warm Worlds and Otherwise*, 1975, p. iv-xviii

[2] 小谷真理. 『女性状無意識: テクノガイネーシス—女性 SF 論序説』勁草書房, 1994.

[3] Kimura, M. “Can a writer disguise the true identity under pen names?: Statistical authorship attribution and the evaluation of variables.”, In *Proceedings of Japanese Association for Digital Humanities 2016*, 2016, p. 16-17.

[4] Burrows, J. F. *Computation into Criticism: A study of Jane Austen‘s novels and an experiment in method*. 1987. Oxford: Clarendon Press.

[5] Burrows, J. F. Not unless you ask nicely: The interpretative nexus between analysis and information. *Literary and Linguistic Computing*, 1992, 7(2), p. 91-109.

[6] Burrows, J. F., & Hassal, A. J. Anna Boleyn and the authenticity of Fielding's feminine narratives. *Eighteenth Century Studies*, 1988, 21, p. 427-453.

[7] Hirst, G. & Feiguina, O. Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, 2007, 22(4), p. 405–417.

[8] Hou, R., and Jiang, M. Analysis on Chinese quantitative

stylistic features based on text mining. *Digital Scholarship in the Humanities*, 2016, 31(2), p. 357-367.

[9] Breiman, L. Random Forests, *Machine Learning*, 2001, 45, p. 5-32.

[10] 金明哲・村上征勝. 「ランダムフォレスト法による文章の書き手の同定」, 『統計数理』, 2007, 55(2), p. 255-268.

[11] Cutler, D. R., Edwards, T. C. Jr., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J. and Lawler, J. J. RANDOM FORESTS FOR CLASSIFICATION IN ECOLOGY. *Ecology*, 2007, 88(11), p. 2783-2792.