機械学習アルゴリズムを用いた防衛白書の計量分析*

Quantitative Analysis of Japan's Defense White Paper Using a Machine Learning Algorithm

河合将志**

Masashi KAWAI**

Abstract

Is there a newly formed concept interpretable as a basis for Japan's security policy that has been becoming prominent since the Gulf War? This study empirically examines this question without focusing on specific concepts such as the "Dynamic Defense Force" and the "Dynamic Joint Defense Force" recently introduced by Japan's Ministry of Defense. By primarily applying random forest, a machine learning algorithm, to the Defense White Papers published by the Ministry between 1970 and 2016, the present author finds it difficult to identify such a concept.

キーワード：防衛白書、計量テキスト分析、機械学習、ランダムフォレスト

Keywords： Defense White Paper, quantitative text analysis, machine learning, random forest

*大阪大学大学院国際公共政策研究科に在学中、竹内俊隆先生から数々のご教示をいただきました。長年のご功労に敬意を表し、改めてお礼を申し上げます。
**大阪大学大学院国際公共政策研究科招聘研究員

1. はじめに

湾岸戦争を機に日本の安全保障政策は積極化しているが、その基盤となる新たな概念は存在するのだろうか。これまでにも「動的防衛力」や「統合機動防衛力」といった近年に政府（防衛省）から打ち出された概念に着目した確認的分析は行われてきたものの、特定の概念を重要なものとして仮定することなく、広範なデータにもとづいてその存否を検証する探索的分析は行われてこなかった[1]。新たな概念が存在するのであれば、湾岸戦争以前のデータと以降のデータとの判別を可能にする特徴を後者の中に見出すことができるはずであり、本稿では計量テキスト分析を用いた防衛白書の探索的分析を通して、そのような特徴の存否を明らかにする。

2. 手法

手法として用いる計量テキスト分析は、一部の内容分析と同様に、語の生起頻度にもとづいてテキストを計量する手法であるが、内容分析では分析者によって事前に指定された特定の語のみがデータとなるのに対し、計量テキスト分析ではテキスト全体やその計量にもとづいて特定した語がデータとなる。したがって、その利用により、探索的分析が可能になるのである。

テキストをデータとすることによって、様々な手法を計量テキスト分析として使用することができるが、本稿ではテキストのカテゴリー判別において高い汎化性能を発揮することで知られるランダムフォレストを主な手法として用いる[2]。ランダムフォレストは集団学習のアルゴリズムの一種であり、弱学習器の決定木は以下のようにして機能する[3]。ノードtにおけるj番目のカテゴリーのデータ$C_j(j = 1, 2, \cdots, J)$の比率を$P(C_j|t)$とすると、ノードtのジニ係数（不純度）$GI(t)$は次式のように定義され、$\Delta GI(t)$を最大化する変数による分岐が行われる[4]。なお、P_pは親ノードt_pにおけるデータの比率を、P_lはt_pから分岐した左側の子ノードt_lにおけるデータの比率を、P_rはt_pから分岐した右側の子ノードt_rにおけるデータの比率を、$GI(t_p)$、$GI(t_l)$、$GI(t_r)$は各ノードのジニ係数を表している。

$$GI(t) = \sum_{j=1}^{J} P(C_j|t)\left(1 - P(C_j|t)\right) = 1 - \sum_{j=1}^{J} P(C_j|t)^2$$

$$\Delta GI(t) = P_p GI(t_p) - \{P_l GI(t_l) + P_r GI(t_r)\}$$

この決定木を弱学習器とするランダムフォレストの仕組みは、以下のように簡約される[5]。

---

[1] Hiroshi Nakanishi, “Reorienting Japan? Security Transformation Under the Second Abe Cabinet,” *Asian Perspective* 39 (2015): 407–410; Toshi Yoshihara, “Japanese Hard Power: Rising to the Challenge,” American Enterprise Institute, August 25, 2014: 5–6, http://www.aei.org/wp-content/uploads/2014/08/-japanese-hard-power-rising-to-the-challenge_152508260874.pdf; 松村昌廣「『動的防衛力』構想の含意と課題」『桃山法学』19 号 (2012): 51–81; 神保謙「新防衛大綱と新たな防衛力の構想」『外交』5 号 (2011): 109–118.

[2] Leo Breiman, “Random forests,” *Machine Learning* 45 (2001): 5–23; 金明哲『テキストデータの統計科学入門』 (東京: 岩波書店, 2009), 194–202.

[3] Leo Breiman, Jerome H. Friedman, Richard A. Olshen and Charles J. Stone, *Classification and Regression Trees* (New York: Chapman & Hall/CRC, 1984); 金『テキストデータの統計科学入門』, 179–186; 河原達也「TVCM 表現要素の消費者反応に対する効果」『行動計量学』43 巻 1 号 (2016): 91.

[4] ランダムフォレストで用いられる不純度としてはジニ係数の他にもエントロピーがあるが、不純度の違いによる分析結果の相違は小さいとされるため、本稿では対数の含まれないジニ係数を使用する。Pang-Ning Tan, Michael Steinbach and Vipin Kumar, *Introduction to Data Mining* (Boston: Addison Wesley, 2005), 150–172.

[5] Trevor Hastie, Robert Tibshirani and Jerome Friedman『統計的学習の基礎: データマイニング・推論・予測』杉山将他監訳（東

1. データからK個のブートストラップサンプル$B_k(k = 1, 2, \cdots, K)$が生成される。
2. B_kから二進分岐かつ未剪定の最大決定木T_kが生成される。分岐にあたっては、ランダムに抽出されたm個の変数のうち、$\Delta GI(t)$を最大化するものが用いられる。
3. カテゴリー判別の場合、未知のデータxに対するT_kの予測値を$\hat{C}_k(x)$とすると、モデルの予測値は多数決$\{\hat{c}_k(x)\}_1^K$によって求められる。
4. OOB（Out-Of-Bag）データと呼ばれるB_kの生成に際してブートストラップサンプリングされることのなかったデータは、T_kに対する未知のデータであることから、それを用いた汎化性能のテストが行われる。

3. データ

既述のように、データとしては防衛省の発行する代表的な刊行物である防衛白書（1970 年版～2016 年版）を用いる[6]。具体的な検証内容については後述するが、新たな概念が存在するのであれば、湾岸戦争以前のカテゴリーのデータ（以前の防衛白書）と以降のカテゴリーのデータ（以降の防衛白書）との判別を可能にする特徴を後者の中に見出すことができるはずであり、70 年代から発行の続く防衛白書をデータとすることにより、この点を検証することができるのである。

分析用のデータ形式であるデータフレームに関しては、要素に正規化頻度を、行ラベルに防衛白書を、列ラベルに正規化頻度の総和にもとづいて抽出した名詞上位 100 語および湾岸戦争「以前」または「以降」というカテゴリーをもつ教師をとるものを作成し、主なデータとして用いる[7]。なお、正規化頻度としているのは、データの長さのばらつきを考慮してのことであり、名詞上位 100 語としているのは、内容に係わる特に重要な語をデータとするためである[8]。

4. 分析

本節では新たな概念の存否を明らかにするため、⑴データのカテゴリー判別、⑵以降のカテゴリーのデータを特徴づける語の抽出、⑶抽出した語の新たな概念としての解釈、それぞれの可否の検証を試みる[9]。

まず、⑴データのカテゴリー判別が可能なのかを検証する。図 1 の横軸はモデルサイズを、上線は以前のカテゴリーの OOB データに対する誤判別率の推移を、中線は OOB データ全体に対する誤判別率の推移を、下線は以降のカテゴリーの OOB データに対する誤判別率の推移を示しており、いず

京: 岩波書店, 2009), 676; 河原「TVCM 表現要素の消費者反応に対する効果」: 91–92.

6 本文・本編をデータとして用いる。防衛省「防衛白書」http://www.mod.go.jp/j/publication/wp/.

7 教師とはモデルの生成と汎化性能のテストに際して、解答として機能する変数のことであり、1970 年版から 1990 年版を「以前」のカテゴリーのデータとして、1991 年版から 2016 年版を「以降」のカテゴリーのデータとしてコーディングした。なお、100 語はクリーニング済みのものである。上記mについては、カテゴリー判別の場合、総語数をnとすると推奨値は$\sqrt{n}$であり、$m = 10$とする。

8 名詞と内容との関係性は強いとされる。鈴木崇史, 影浦峡「名詞の分布特徴量を用いた政治テキストの分析」『行動計量学』38 巻 1 号 (2011): 83–92.

9 ソフトウェアとしては R（Ver. 3.3.1）を、ランダムフォレストを実装するためのパッケージとしては randomForest を、日本語形態素解析器 MeCab を実装するためのパッケージとしては RMeCab を用いる。R Core Team, *R: A Language and Environment for Statistical Computing* (Vienna: R Foundation for Statistical Computing, 2016), https://www.R-project.org/; Andy Liaw and Matthew Wiener, "Classification and Regression by randomForest," *R News* 2, no. 3 (2002): 18–22, http://CRAN.R-project.org/doc/Rnews/; Taku Kudo, Kaoru Yamamoto and Yuji Matsumoto, "Applying Conditional Random Fields to Japanese Morphological Analysis," *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)* (2004): 230–237; Taku Kudo, "MeCab: Yet Another Part-of-Speech and Morphological Analyzer," http://taku910.github.io/mecab/; Motohiro Ishida, "RMeCab: interface to MeCab R package version 0.99994," 2016, https://sites.google.com/site/rmecab/.

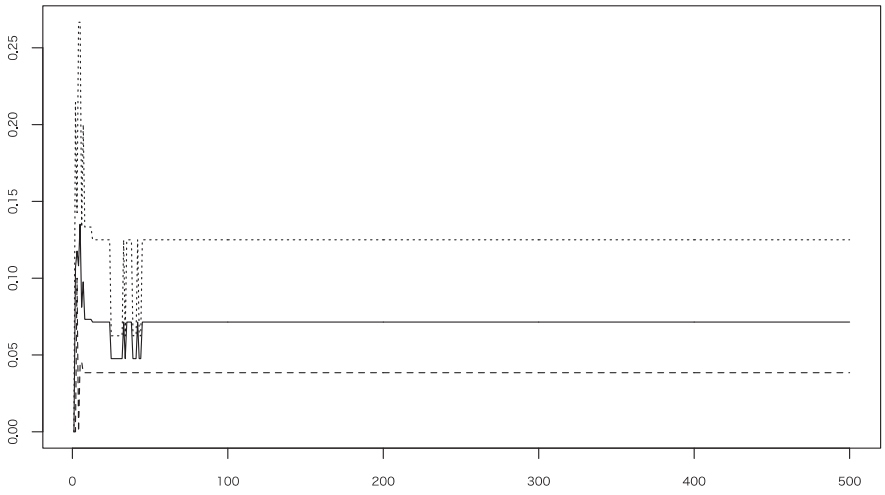

図 1　誤判別率の推移

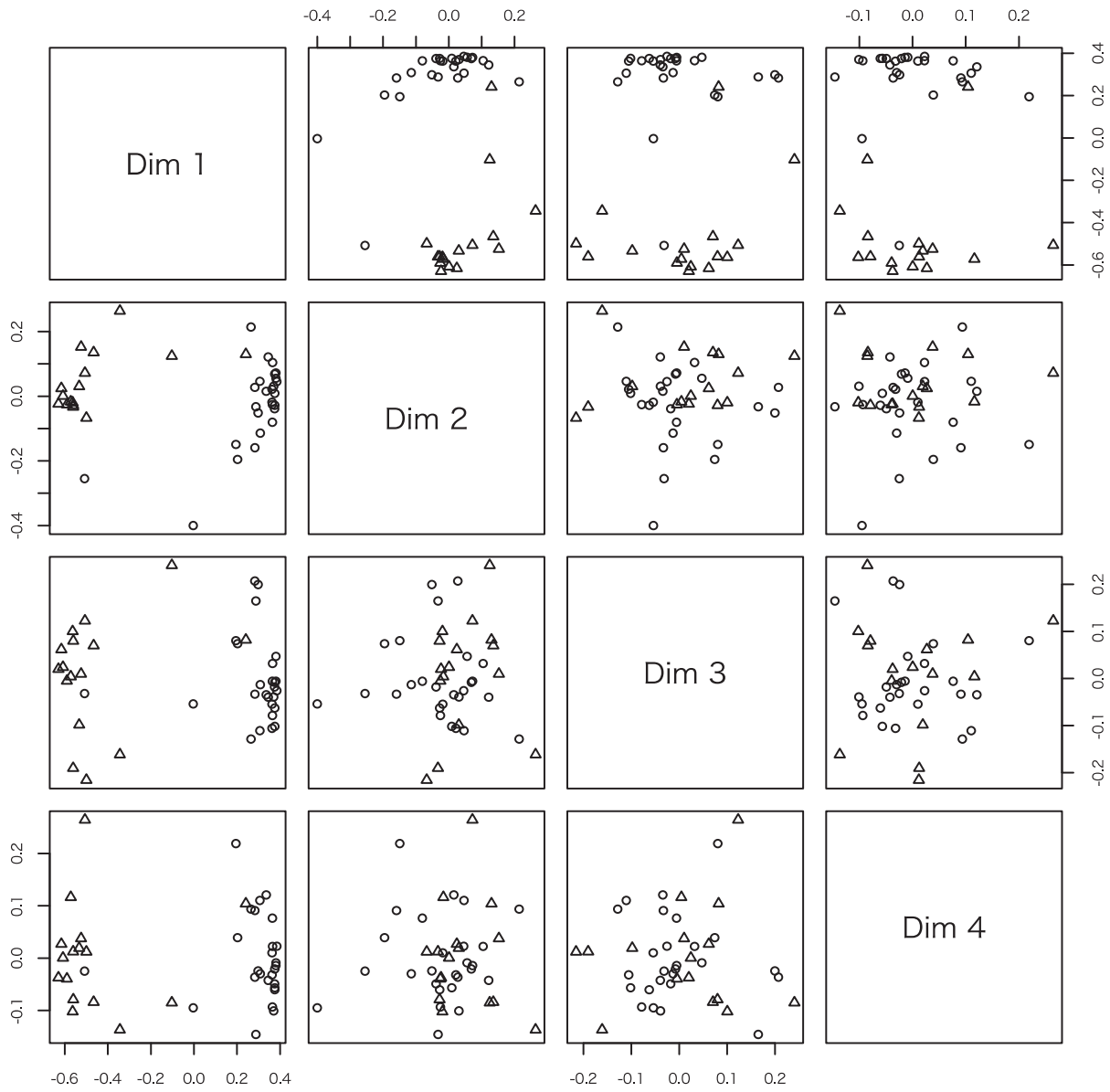

図 2　MDS プロット

れの誤判別率も安定して低く推移していることが確認できる。また、ランダムフォレストの分析結果をもとに作成した図 2 の MDS（Multi-Dimensional Scaling）プロットからは、以前のカテゴリーとされるデータ（△）と以降のカテゴリーとされるデータ（○）がわかれて布置される傾向がうかがえ、これらの分析結果を踏まえれば、カテゴリー判別は可能と言える。

続いて、⑵以降のカテゴリーのデータを特徴づける語の抽出ができるのかを検証する。図 3 は判別への寄与度（ジニ係数の減少量の平均値）にもとづいて抽出した上位 10 語の値であり、図 4 の横軸は正規化頻度を、実線は以前のカテゴリーについての部分従属度（partial dependence）の推移を、点線は以降のカテゴリーについての部分従属度の推移を示している。図からは、「航空」、「能力」、「艦」、「戦闘」、「向上」、「力」、「侵略」、「国民」、「整備」の正規化頻度と以前のカテゴリーについての部分従属度との関係が正比例する傾向にあり、残る「活動」の正規化頻度と以降のカテゴリーについての部分従属度との関係が正比例する傾向にあることがわかる。つまり、「活動」が以降のカテゴリーのデータを特徴づける語であるということが示されているのである。

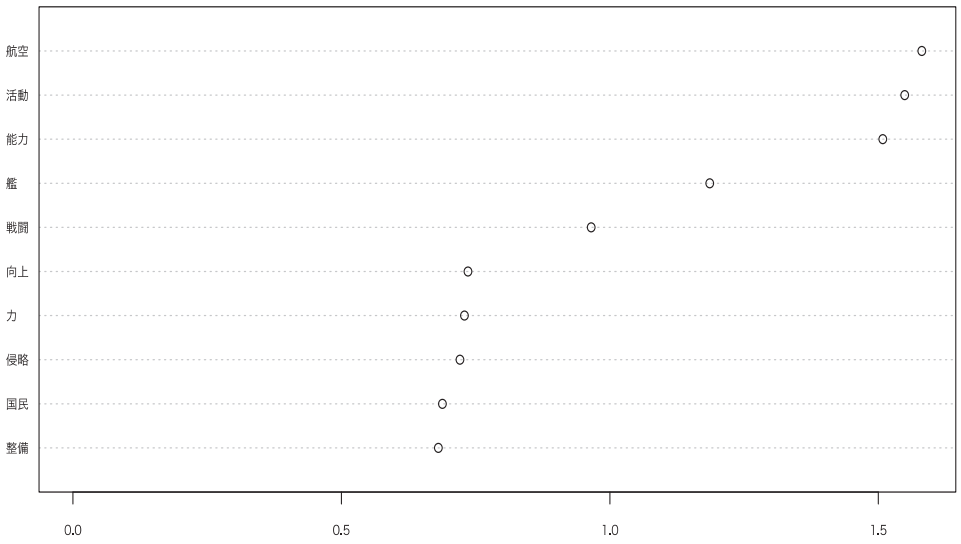

図 3　寄与度グラフ

最後に、⑶「活動」を新たな概念として解釈できるのかを検証する。図 5 は「活動」の意味を正確に把握するために、その共起情報から作成したネットワークマップである[10]。図からは、「活動」が「情報収集」、「警戒監視」、「偵察」、「調整」、「相互」、「措置」、「実施」、「協力」、「共同」、「isr」、「intelligence」、「surveillance」、「reconnaissance」、「pko」と共起しており、それが主に情報収集・警戒監視・偵察（Intelligence, Surveillance and Reconnaissance: ISR）活動や平和維持活動（Peace Keeping Operation: PKO）に係る語であることがわかる。以降のカテゴリーのデータが ISR 活動や PKO に係る語である「活動」によって特徴づけられることに不思議はないが、主に兵器や兵力に係る語として括ることのできる「航空」～「整備」の正規化頻度の総和の推移（上線）と、「活動」の正規化頻度の推移（下線）を示した図 6 を見る限り、「活動」を新たな概念として解釈するのは難しい[11]。「活動」

[10] 共起情報を取得するための関数としては三宅真紀氏が開発したもの（非公開）を用い、共起頻度 4 以上の語をクリーニングした上でデータとした。 また、描画パッケージとしては igraph を使用した。Gabor Csardi and Tamas Nepusz, “The igraph software package for complex network research,” *InterJournal* Complex Systems (2006): 1695, http://igraph.org.

[11] データ形式変換パッケージとしては reshape を、描画パッケージとしては ggplot2 を用いた。Hadley Wickham, “Reshaping data with the reshape package,” *Journal of Statistical Software* 21, no. 12 (2007): 1–20; Hadley Wickham, *ggplot2: Elegant Graphics for Data Analysis* (New York: Springer-Verlag, 2009).

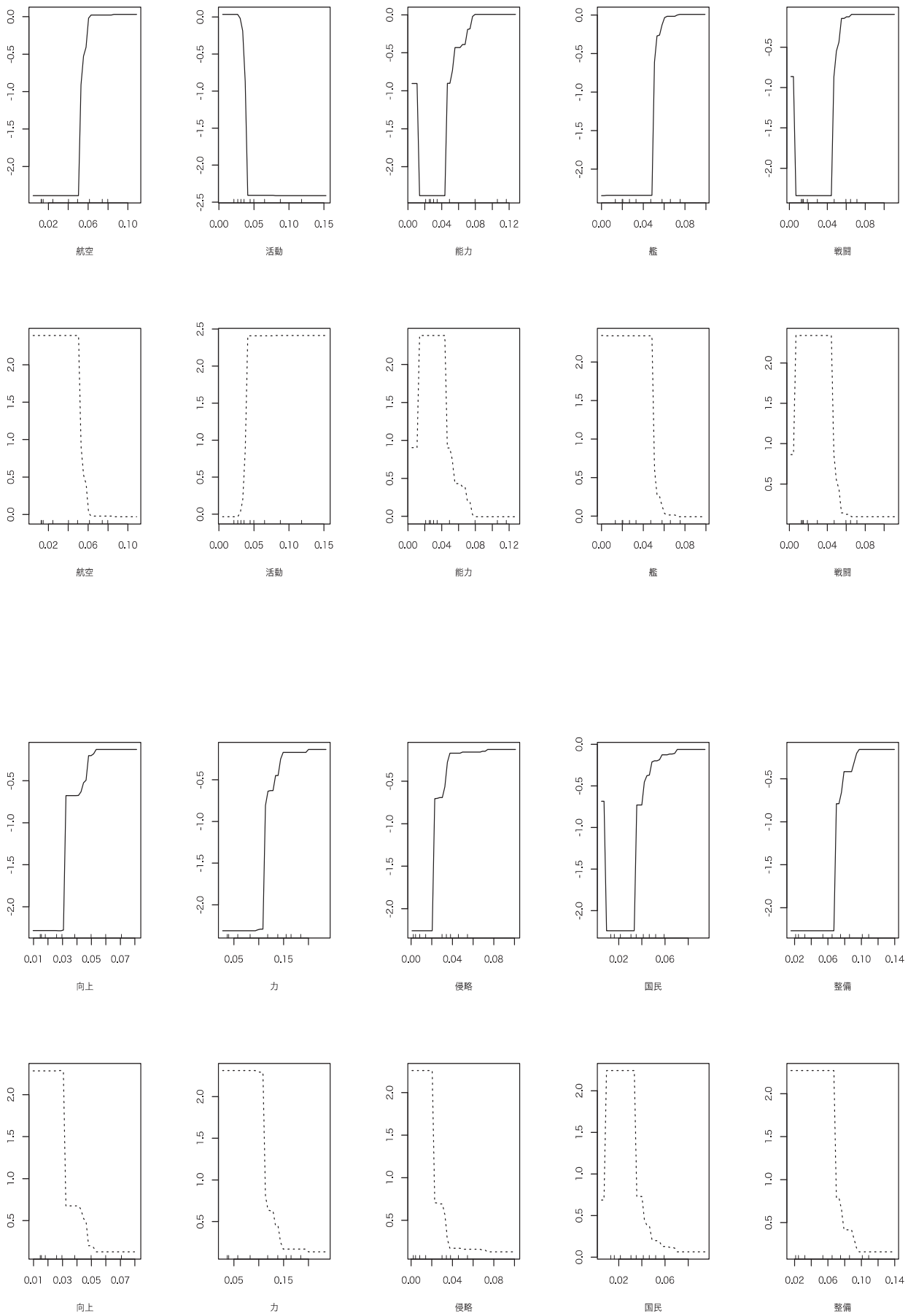

図 4　部分従属度の推移

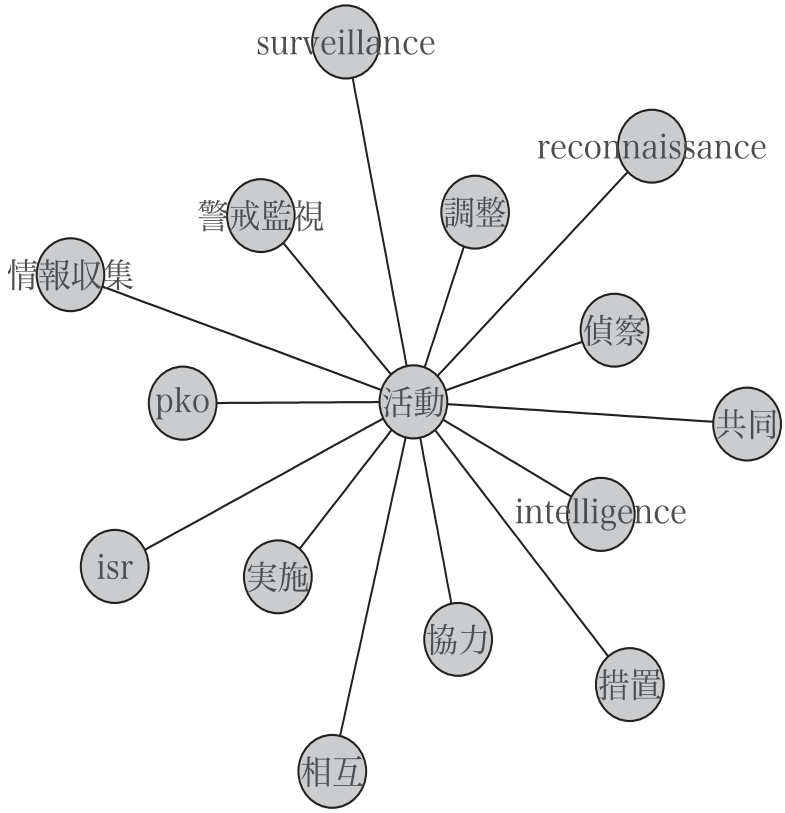

図 5　ネットワークマップ

図 6　正規化頻度の推移

が唯一の語であることから予想されるように、「航空」～「整備」の正規化頻度と比べ、その正規化頻度は以降のカテゴリーのデータにおいてさえ低く、以降のカテゴリーのデータが強く特徴づけられているとまでは言えないからである。

5. おわりに

本稿ではランダムフォレストを主に用いた防衛白書の分析を通して、湾岸戦争以降の安全保障政策の基盤となる新たな概念の存否を検証し、その存在に否定的な見解を示した。既存研究においては重要なものとして仮定された特定の概念に着目した確認的分析が行われてきたのに対し、本稿では特段の仮定を設けることなく、広範なデータにもとづいてその存否を検証する探索的分析を行い、湾岸戦争以降の防衛白書はそれほど積極的には特徴づけられないということを実証したのである。新たな概念が存在するのであれば、湾岸戦争以前の防衛白書と以降の防衛白書との判別を可能にする特徴を後者の中に見出すことができるはずであるが、その多くを前者の中に特定したのであった。

安全保障政策の指針としてだけではなく、それに対する国民の理解を促すものとしても機能するであろう明確な概念の存在は、その積極化にあたり、都合の良いもののように思われるにも関わらず、なぜこのような分析結果がえられたのだろうか。柔軟な政策決定を可能にする環境が整ったことなどがその理由として考えられるが、この点の解明については今後の課題としたい。