JAIRO Cloud 利用機関におけるグリーンオープンアクセス進捗度に関する予備的分析

河合　将志[1)], 林　正治[1)], 新妻　聡[1)], 尾城　孝一[1)], 西澤　正己[1)], 山地　一禎[1)]

1) 国立情報学研究所

m-kawai@nii.ac.jp

Preliminary Analysis on the Progress of Green Open Access among JAIRO Cloud User Institutions

Masashi Kawai[1)], Masaharu Hayashi[1)], Akira Niitsuma[1)], Koichi Ojiro[1)], Masaki Nishizawa[1)], Kazutsuna Yamaji[1)]

1) National Institute of Informatics

概要

クラウド型の IR（Institutional Repository）環境提供サービスである JAIRO Cloud は、多くの機関に導入され、日本を代表するリポジトリサービスとなりつつあるが、学術雑誌論文の登録件数には利用機関の間で大きな差が見られる。本研究では、この差を生み出している要因を明らかにするため、利用機関に対してアンケート調査を行い、その結果を主なデータとして計量分析を行う。そして、この差が図書館の業務に係る変数ではなく、機関の研究力に係る変数などによるものであることを示す。

1　はじめに

ここ 20 年ほどの間、査読付きの学術雑誌論文の OA（Open Access）化に向けた取組みが、近年のオープンサイエンス推進の潮流と相まって、世界的に広まっている。

OA を実現するための方法としては、学術雑誌自体を OA 雑誌として出版する方法（ゴールド OA 方式）と、学術雑誌論文を誰もが無料でアクセスできるリポジトリに登録して公開する方法（グリーン OA 方式）との二つが提唱されている。最近では、ゴールド OA への転換を大規模に進めようとする OA2020 のようなイニシャティブが主流となっているが［1］、過去に出版された学術雑誌論文には、ゴールド OA 方式を適用できないという課題もあり、グリーン OA は OA 戦略のなかで依然として重要な意味をもつ。

このグリーン OA の受け皿となる OA リポジトリには、大学等の機関が設置する IR と分野別のリポジトリとがあるが、日本の IR 設置数は 754（2018 年 3 月現在）に達しており、世界でも有数の IR 保有国となっている［2］。しかしながら、日本の機関に所属する研究者による学術雑誌論文に占めるグリーン OA の割合は、世界平均（10.5%）を下回る 9.9%に留まっており［3］、日本の IR が OA の推進に十分に活用されているとは言い難い。

こうした状況を改善し、日本のグリーン OA を推し進めるためには、日本の IR の 66%を占める JAIRO Cloud を活用した取組みが不可欠であり、この取組を検討するための一助として、本研究では計量分析を行う。

具体的には、JAIRO Cloud 利用機関の間で見られる学術雑誌論文の登録件数の差に着目し［4］、機械学習アルゴリズムを用いて、登録に積極的な機関と残る機関との判別タスクを行うとともに、判別に寄与した変数の特定を試みる。

2　データ

データには、従属変数（教師）として「学術雑誌論文登録件数ダミー」を、説明変数として後述の変数を、個体としてこれらについての情報が取得可能な JAIRO Cloud 利用 308 機関をとるデータフレームを使用する。

従属変数の「学術雑誌論文登録件数ダミー」は、全登録件数の約 9 割を占める一部の積極的な機関と残る機関とを分ける変数である。両者を区別するため、値として「学術雑誌論文登録件数 1000 件以上」と「以下」とを設け、前者を 1 と、後者を 0 とコーディングした。

説明変数には以下に挙げる 22 の変数を用いる。

これらのうち、図書館の業務に係る変数である 2, 4–7, 10, 12, 14–22 のコーディングは、アンケート調査を実施し、その結果にもとづいて行った。ダミー変数のコーディングについては、「有り」、「国公立（設置種別）」、「該当」、「学術雑誌論文産出件数 1000 件以上」、「参加（DRF: Digital Repository Federation）」を 1 と、「無し」、「私立」、「非該当」、「1000 件以下」、「不参加」を 0 とした。

1. 科研費件数［5］［6］
2. 学術雑誌論文提供依頼ダミー
3. 設置種別ダミー
4. ダウンロード件数通知ダミー
5. 登録者_著者自身ダミー
6. 登録者_図書館員・事務職員ダミー
7. 登録者_両者ダミー
8. 博士授与件数［7］
9. 学術雑誌論文産出件数ダミー［8］
10. APC 支援ダミー
11. DRF ダミー［9］
12. IR 委員会ダミー
13. IR 運用期間［4］
14. IR 業務外部委託ダミー
15. IR 業務規定ダミー
16. IR 専用定常予算ダミー
17. IR 担当職員数
18. OA 委員会ダミー
19. OA ウィークイベントダミー
20. OA 広報資料ダミー
21. OA 説明会ダミー
22. OA 方針ダミー

個体については、「学術雑誌論文登録件数 1000 件以上」の機関数と「以下」の機関数との不均衡に起因する判別精度の低下を避けるため、SMOTE（Synthetic Minority Over-sampling Technique）による前処理を施した[5]。これにより、「登録件数 1000 件以上 21 機関：以下 287 機関」であった処理前の個体数を、「登録件数 1000 件以上 42 機関：以下 42 機関」へと均衡化した［10］。

3　手法

手法には、機械学習アルゴリズムであるランダムフォレストを用いる［11-13］。ランダムフォレストを使用することにより、判別に寄与した変数の特定を、判別の精度を維持しつつ行うことできるからである。ランダムフォレストの他にも、サポートベクターマシンや深層学習などが精度の高いアルゴリズムとして知られているが、それらのアルゴリズムでは寄与した変数を特定することはできず、本研究の目的を達成するための手法としては適当ではない。

以上のような特徴をもつ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［14, 15］。

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、ノードの分割基準である不純度の減少量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、それを用いた交差検証に相当するテストが行われる。

判別に寄与した変数の特定は、不純度の減少量の平均値にもとづいて行うことができ、各変数がどのように判別に寄与したのかは、変数の値と部分従属度との関係をプロットした部分従属図を作成することによって把握できる［16］。

なお、ハイパーパラメーターであるモデルサイズと上記mについては、それぞれ 10000 と 5 に設定し、分割基準である不純度については、ジニ係数を用いる。

4　分析結果

表 1 は DRF ダミーOOB データを用いたテストの結果であり、正答（39, 39）が誤答（3, 3）を大きく上回っている。

表 1 混同行列

	登録件数1000件以上（予測）	登録件数1000件以下（予測）
登録件数1000件以上（実際）	39	3
登録件数1000件以下（実際）	3	39

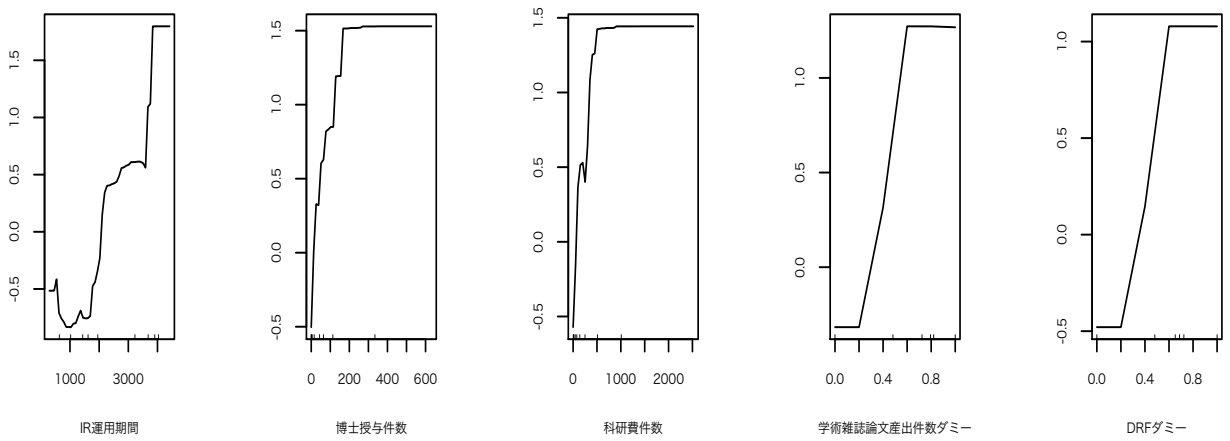

図 2 部分従属図

図 1 の横軸は判別への寄与度をあらわすジニ係数の減少量の平均値である。表には「IR 運用期間」～「DRF ダミー」の寄与度が大きいことが示されている。

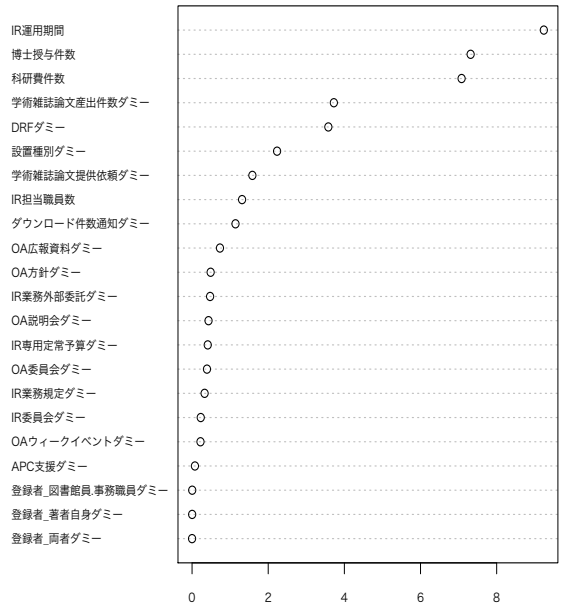

図 1 寄与度グラフ

図 2 は「IR 運用期間」～「DRF ダミー」の値（横軸）に対する、カテゴリー「登録件数 1000 件以上」についての部分従属度（縦軸）をプロットしたものであり、プロットは全て右肩上がりになっている。つまり、これらの変数の値が大きい個体を、モデルは登録件数 1000 件以上の機関として判断する可能性が高いのである。

5 おわりに

以上のように、グリーン OA 学術雑誌論文の登録件数を左右する主な変数は、「IR 運用期間」、「博士授与件数」、「科研費件数」、「学術雑誌論文産出件数ダミー」、「DRF ダミー」であることが明らかになった。登録件数が 1000 件以上の積極的な機関は、これらの値が大きい傾向にあるのである。「博士授与件数」、「科研費件数」、「学術雑誌論文産出件数ダミー」は、機関の研究力に係るものであることから、登録に積極的な機関は、研究に注力する傾向にあると言うこともできよう。

「IR 運用期間」～「DRF ダミー」の寄与度が大きいことそのものに不自然はないものの、図書館の業務に係る多くの変数の寄与度が総じて低く留まっていることは、直感に反するものであり、今後は変数のコーディング法を変えるなどして、寄与度の異同を確認する必要があるだろう。

参考文献

[1] Budapest Open Access Initiative, Read the Budapest Open Access Initiative, http://www.budapestopenaccessinitiative.org/read.

[2] 国立情報学研究所、機関リポジトリ公開数とコンテンツ数の推移、https://www.nii.ac.jp/irp/archive/statistic/.

[3] Martín-Martín, A., Costas, R., van Leeuwen, T. and López-Cózar, E. D., Evidence of Open Access of Scientific Publications in Google Scholar: A Large-Scale Analysis, Journal of Informetrics, vol. 12, no. 3, pp. 819–841, 2018.

[4] 国立情報学研究所、IRDB コンテンツ分析（2018 年 3 月データ収集）、http://irdb.nii.ac.jp/analysis/browse.php.

[5] 文部科学省、平成 25 年度科研費（補助金分・基金分）の配分について、http://www.mext.go.jp/a_menu/shinkou/hojyo/__icsFiles/afieldfile/2013/05/20/1335064_01.pdf.

[6] 国立情報学研究所、KAKEN、https://kaken.nii.

ac.jp/ja/index/.
[7] 文部科学省、平成 25 年度博士・修士・専門職学位の学位授与状況、http://www.mext.go.jp/component/a_menu/education/detail/__icsFiles/afieldfile/2017/01/26/1299723_10.pdf.
[8] 科学技術・学術政策研究所、研究論文に着目した日本の大学ベンチマーキング 2015、http://www.nistep.go.jp/wp/wp-content/uploads/NISTEP-RM243-MaterialJ01.pdf.
[9] DRF、参加機関一覧、http://drf.lib.hokudai.ac.jp/drf/index.php?参加機関一覧.
[10] Chawla, N., Bowyer, K., Hall, L. and Kegelmeyer, W. P., SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
[11] Breiman, L., Random Forests, Machine Learning, vol. 45, pp. 5–23, 2001.
[12] Liaw., A. and Wiener, M, Classification and Regression by Randomforest, R News, vol. 2, no. 3, pp. 18–22, 2002, http://CRAN.R-project.org/doc/Rnews/.
[13] R Core Team, R: A Language and Environment for Statistical Computing, https://www.R-project.org/.
[14] Hastie, T., Tibshirani, R. and Friedman, J., 杉山将他監訳、統計的学習の基礎: データマイニング・推論・予測、676 頁、岩波書店、2009 年.
[15] 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、91 頁、2016 年.
[16] Friedman J., Greedy Function Approximation: A Gradient Boosting Machine, The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.