情報誌「大学評価と IR」第 12 号　令和 3 年（2021 年）9 月
［事例報告］（大学評価コンソーシアム）

科学研究費申請種目の選択を支援する分類器の作成

岩田 博夫[1]

概要：日本学術振興会が所掌する科学研究費への申請支援は、リサーチ・アドミニストレーター(URA)の重要な業務の一である。URA が研究者に適した申請種目を助言するときに用いるクラス分類器を機械学習を用いて作成した。特徴量として研究者の研究機関、職位、男女別、研究分野、科学研究費初受給年と申請時年までの発表論文数を用いた。URA はこの分類器を用いることで、軋轢が少なく研究者に適した申請種目を指し示すことが出来るようになり、さらに、研究種目のステップアップの助言も行えるようになると考える。

キーワード：科学研究費、基盤研究(A)、 基盤研究(B)、基盤研究（C)、課題代表者、機械学習、クラス分類器

１．はじめに

リサーチ・アドミニストレーター(URA)の業務の一つとして日本学術振興会が所掌する科学研究費への申請支援がある。URA から研究者に「先生は実績がおありですので、今年はステップアップして基盤(A)に応募しましょう。」と勧めたときに、研究者から「落ちたら研究が止まってしまう。基盤(B)で手堅くいきたい。」とか、URA から若手の助教に「基盤(A)とは余りにも大胆な！今年は基盤(C)に申請しましょう。」などのやり取りがあると想像される。ときには場の雰囲気がかなり険悪になり、URA の方も気まずい思いをされることもあるであろう。また、研究費に間接経費が付いているので、大学の執行部は研究者に出来るかぎり高額の研究費を獲得してほしい。この様な状況から、URA が研究者に適切な、また、出来ればより高額の研究費が交付される種目への応募を助言できるようにする試みが行われ、論文としても発表されてきた（久保・伊藤, 2020)。本稿では、URA がより直接的に研究者に適切な研究種目を指し示すことを支援する方法を提案する。この様な方法があれば、話し合いがよりスムーズに進むと考える。

本稿では科学研究費助成事業の中心事業である基盤研究(A)、 (B)、 (C)（以降、基盤 A、 B、 C と略記）を対象とした。科学研究費に採択された各種目約 500 名、計約 1500 名の課題代表者（以降、代表者）の申請時までの論文発表数、所属大学、職位、男女等の公開情報を特徴量として機械学習により種目別クラス分類器（以降、分類器）の作成を行い、この分類器に個々の研究者の特徴量を入力すれば、その研究者に適した研究種目が表示されるようにする。

[1] 京都大学 COI 拠点研究推進機構 機構戦略支援統括部門 部門長　メール：hiwata00@gmail.com

２．方法

2.1. 検討対象者

対象は 2014 年に採択された基盤 A、B、C の代表者である（KAKEN: 科学研究費助成事業データベース[2]）。基盤 A では全課題の代表者を対象とし、基盤 B と C では全代表者からランダムに約 600 課題の代表者を選び、その中から、法学、経済学また人文系研究者等は、発表論文数データを Scopus[3]から集めることが困難であったので、今回は検討対象から外した。検討対象は、それぞれ基盤 A: 504 名、基盤 B: 440 名、基盤 C: 496 名の代表者とした。

2.2. 機械学習

機械学習アルゴリズムは、データのスケール変換が必要でなく、また、デフォルトのパラメータで十分機能するランダムフォレスト（Müller and Guido, 2017）を採用した。検討対象者をランダムにほぼ 75:25 になるように訓練群と検証群に分けた。すなわち基盤 A: 378 名、基盤 B: 344 名、基盤 C: 360 名を訓練群とし、それぞれの残り基盤 A: 126 名、基盤 B: 96 名、基盤 C: 136 名を検証群として用いた。ターゲットデータとして、訓練群の基盤 A: 378 名にラベル A、基盤 B: 344 名にラベル B、基盤 C: 360 名にラベル C を付した。訓練群を教師データとして機械学習を行い、検証群の研究者を基盤 A, B, C の種目に割り振る分類器を作成した。

採用した特徴量の研究機関、職位、男女、研究分野、科学研究費初受給年は科学研究費助成事業データベースから、各人の 2013 年までの総論文数と 2008~2013 年の各年の論文数は Scopus から得た。研究機関は、国公立研究機関: 研究機関、私立大学：私大、公立大学：公大、国立大学にわけ、さらに国立大学は第３期中期目標期間において採用された重点支援別に国①、国②と国③の３類型に分けた（例えば、文部科学省, 2015）。また、工業高等専門学校は件数も少ないこともあり国①に分類し、民間企業等はその他とした。研究分野としては、申請時に記載された系｜分野｜分科｜細目名の上位から二つ目の分野を用いた。研究機関、職位、男女、研究分野のカテゴリー変数は、one-hot-encoding（Müller and Guido, 2017［日本語訳版］207 ページ）により 0 と 1 へ数値化し、機械学習を行いやすくした。2013 年までの総論文数、各年の論文数、科学研究費初受給年はスケーリングを行わずにそのまま量的変数として採用した。

表１．代表者の科学研費初受給年と 2013 年までの総論文数

		科学研究費初受給年	総論文数
全体	平均	1998.9	96.1
	標準偏差	8.1	120.1
基盤A	平均	1995.0	149.7
	標準偏差	7.3	121.9
基盤B	平均	1998.9	86.0
	標準偏差	7.4	81.6
基盤C	平均	2002.7	50.7
	標準偏差	7.6	125.6

[2] https://nrid.nii.ac.jp/ja/index/
[3] Search for an author profile - Scopus: https://www.scopus.com/freelookup/form/author.uri

３．結果

3.1. 種目別特徴

本項で解析対象とした代表者の特性を簡単に述べる。いずれの研究種目においても代表者の職位としては教授職が一番多いものの、その比率は基盤 C → B → A の順に増加している。各研究種目で研究期間３～５年間に申請できる総研究費は、基盤 C: 500 万円以下、 基盤 B: 2,000 万円以下、基盤 A: 5,000 万円以下である。研究費が大きくなるに従い、申請主体が若手助教クラスからより成熟し、また、教室運営責任者である教授クラスに移っていくことを反映しているのであろう。採択された課題の研究分野別では、医歯薬学が基盤 C では実に約 44%を占め、基盤 B でも依然として分野別１位であるが、基盤 A では３位に後退した。

代表者の科学研究費初受給年と総論文数を種目別に表１にまとめた。科学研究費初受給年は A<B<C、総論文数は基盤 C<B<A の順に値が大きくなっている。若手研究者中心の基盤 C からより研究歴が長く成熟した研究者へと代表者が移って行っていることがわかる。平均値から見る限りでは、基盤 A、B、C は対象者の研究歴、すなわちジュニア、中堅、シニアに応じた研究種目別によく制度設計されているように見える。しかし、総論文数の標準偏差が大きいことから見て取れるように種目間でかなりの重なりが見られる。

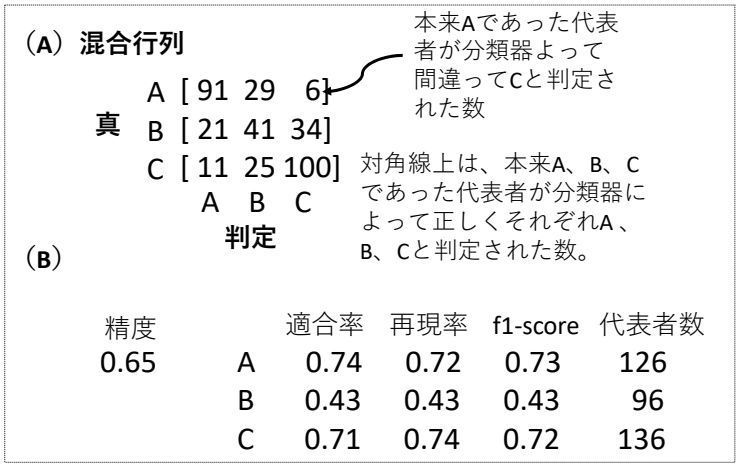
3.2. 分類器

3.2.1. 基盤 A、B、C のクラス分類

訓練群データから機械学習することで作成した分類器を用いて検証データ基盤 A: 126 名、基盤 B: 96 名、基盤 C: 136 名の３クラス分類を行った結果を図１に示した。また、図中には混合行列の読み方も簡単に書いてある。混合行列で、左から“真”→“A”→ “91”→ “29” →“6”とたどる経路の数字の意味は、基盤 A に採択された 126 名の代表者が、試作した分類器により 91 名が正しく基盤 A の代表者として分類され、残りの 35 名うち 29 名が基盤 B に、 6 名が基盤 C へと分類されたことを示している。混合行列の下から“判定”→ “A” →“11” → “21” → “91”とたどる経路は、分類器により基盤 A と判定された 123 名の代表者のうち、本来基盤 C、B の代表者である者がそれぞれ 11 名と 21 名が間違って基盤 A 代表者と判定され、正しく基盤 A と判定された代表者は 91 名であることを示す。混合行列の対角線の数字”91”、 “41”、 “105”は本来の課題に正しく分類された人数を示す。本来の種目以

図１：A,B,C の３クラス分類

(A)：混合行列とその読み方　(b)：適合率や再現性



外に分類された代表者が多数おられ、特に基盤Bの代表者は、本来の基盤Bに分類された代表者は41名と一番多いものの、基盤Aと基盤Cに分類された代表者数の合計は、基盤Bに正しく分類された41名を上回る55名にもなっている。種目間で特徴量の重なりが大きいことが影響しているのであろう。

混合行列より算出でき、より簡便に全体像を把握できる指標として適合率(precision)と再現率(recall)がある。基盤Aの適合率（Aと判定された者が実際にAである割合）は、図1の混合行列から、

適合率 ={ 91/(91 + 21 + 11)} = 0.74

と計算される。また、基盤Aの再現率（実際に基盤Aの代表者である者うちAと判定された割合）は、

再現率 = {91/(91 + 29 + 6)} = 0.72

と計算される。

適合率と再現率の2つを1つにまとめる評価指標として、

f_1 = 2(適合率 x 再現率)/ (適合率 ＋ 再現率)で定義されるf1-scoreがある。図1には f1-scoreも示した。

より簡便に分類の良さを見るのによく使われ精度(Score）とは正しく判定された代表者数をすべての代表者総数で割ったものであり図1に示した例では、

精度=(91+41+100)/(126+96+136)= 0.65

となる。

3.2.2. 特徴量

分類器は特徴量を用いて代表者を分類し、混合行列を算出する。図2にはランダムフォレストが図1のクラス分類を行うときに重要視した特徴量の程度を示した。総論文数が一番重要視されている。総論文数が多いということは、長年たゆまずに研究し（長い研究歴)、着実に論文が採択され（興味深い研究)、さらに、良い共同研究者に恵まれて彼ら（彼女ら）の論文の共著者になる等、多くの情報が

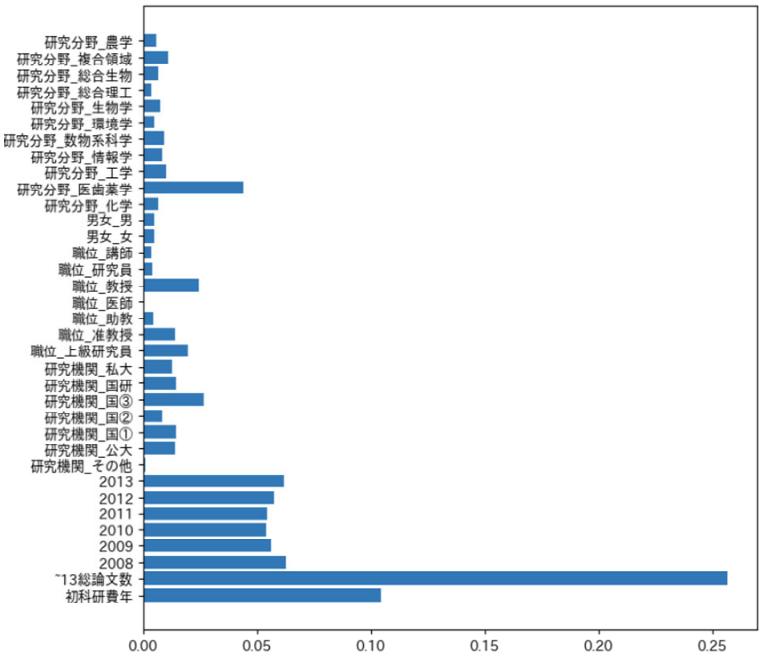


図２：代表者をA, B, Cへと分類する時の特徴量の重要度

一つにまとめられた特徴量であるからであろう。年次論文数(2008~2013)の重要度の変動はあまり見られなかった。他の特徴量で突出しているのは研究分野の医歯薬学である。これはおそらく医歯薬学分野の代表者が
基盤 C では実に 44%を占めていることが反映されているためであろう。他のわずかであるが重視された特徴量は、国③（第 3 期中期目標に示された重点支援③に分類された大学）と職位教授であろう。基盤 A の 505 名の代表者を見ると、職位が教授である者が 439 名で、また、315 名の代表者が国③に所属していることが反映されているのであろう。

3.2.3. ２クラス分類

基盤 A、B、基盤 A、C、また、基盤 B、C 間の２クラス分類を行い、その結果を表２に示した。当然予想されることではあるが基盤 A と C 間は良く分類でき、精度は 0.90 であった。一方、基盤 A、B 間と基盤 B、C 間の分類の精度は 0.72 と 0.71 とまずまずの分類結果であった。基盤 B の 108 名の代表者のうち、36 名も基盤 A に分類され、再現率が 0.67 と低くなっている。基盤 A、B 間また B、C 間の２クラス分類の分離の悪さは、機械学習に用いたアルゴリズム・ランダムフォレストの性能が低いのではなく、基盤 B が制度上中堅研究者向けであり、基盤 B の代表者には基盤 A と、または基盤 C との境界域におられる研究者が多いことが原因だと思われる。

表２：基盤 A,B,C 間の課題代表者の２クラス分類

	混合行列	精度		適合率	再現率	f1-score
AB分類	[97 31] [36 72]	0.72	A B	0.73 0.70	0.76 0.67	0.74 0.68
AC分類	[109 14] [10 117]	0.90	A C	0.92 0.89	0.89 0.92	0.90 0.91
BC分類	[75 34] [33 93]	0.71	B C	0.69 0.73	0.69 0.74	0.69 0.74

3.2.3. 基盤 B、 C 代表者のステップアップ

2014 年に基盤 B と C に採択された代表者をランダムにそれぞれ 111 名と 106 名を選び、その後の基盤研究費の採択状況を調べた。その結果を表 3 にまとめた。基盤研究では研究期間が 3~5 年とされているが、多くの研究者は 3 年で終了し次の科研費に応募していた。すなわち 2017~2021 年の期間に 2 回採択される可能性がある。実際に 2 回採択された研究者も見られた。表 3 中の採択無しは、応募されたか否かはわからないが、2017 年度以降に基盤研究に代表者として採択履歴がない方の数である。現状維持は同じ研究種目に採択された方の数である。ステップアップは基盤 B の代表者が 2 回の応募のどちらかで基盤 A に採択された代表者数、基盤 C の代表者が 2 回の応募のどちらかで基盤 B に採択された代表者数である。今回調べた

表３：基盤 B, C 代表者のその後の採択

	採択無し	現状維持	ステップアップ
B: 111	23(21%)	69(62%)	19 (17%)
C: 106	33(31%)	56(53%)	17 (16%)

基盤 C の 106 名の代表者でこの期間の間に基盤 A へステップアップした代表はいなかった。

基盤 B の代表者 111 名で、2 回採択される可能性があったにも関わらずわずか 17%の 19 名しかステップアップを果たせていない。同様なことは基盤 C の 106 名代表者についても言え 16%の 17 名しかステップアップを果たせていない。同じことを言うことになるが、2017~2021 年の 5 年間に基盤研究の代表者として採択されなかった、また、以前の同一種目への採択であり、ステップアップが見られなかった代表は、基盤 B で 83%、基盤 C で 84%にもなる。研究費獲得の確実性を重視しているのか、現状維持志向の強さが目を引く結果になっている。

3.2.4. 分類器を用いた助言

URA が今回作成した分類器を使用する状況を考える。これから面談しようとする研究者の特徴量、すなわち研究機関名から過去 6 年間の各年の発表論文数までを集めて分類器に入力し分類する。例として 2014 年度の採択課題の代表者 3 名の分類を行った結果を表 4 に示した。実績カラムは、今回 2014 年度に実際に採択された種目を示してある。これから申請しようとする研究者の分析では、このカラムは空白になる。確率カラムの数字は分類器が算出した各代表者の基盤 A、B、C への適合確率である。

表 4 ：研究者への助言に用いるデータシート

研究者名	研究機関	初科研費	論文数		予想確率				実績
		受給年	2008	2009	2013	A	B	C	
(イ)	国③	1988	10	11	15	0.98	0.02	0.00	A
(ロ)	国②	2014	0	1	4	0.00	0.01	0.99	C
(ハ)	国①	1993	4	1	4	0.73	0.22	0.05	B

研究者（イ）の場合は、2013 年までに取得可能なデータから分類器は A 確率として 0.98 と極めて高い値を出している。URA は迷わずに研究者（イ）に「基盤 A へ申請するのが適切である」と助言できる。実績欄に示すように、実際に 2014 年度に基盤 A に採択されている。研究者（ロ）では、C 確率が 0.99 と高く、この場合も「基盤 C へ申請するのが適切である」と助言でき、実際に基盤 C に採択されている。研究者（ハ）の場合は、A 確率は 0.73 で B 確率は 0.21 であるので、分類器は基盤 A に申請するのが適切であると判定している。しかし、この研究者は基盤 B へ申請して採択されている。この研究者（ハ）の場合が URA の力の見せどころである。

科研費申請の採否判断はヒトの査読者により行われる。査読時に最も重要視されるのは研究内容である。一方、上記の分類器には研究内容に関連する特徴量をほとんど含んでいない。URA は研究者の申請書案をじっくり読んで、やはり研究者（ハ）が基盤 A に申請するのが適当と考えるのであれば、表 4 のデータシートを示しながら「過去のデータからも、先生の実績であれば基盤 A 採択される可能性は高いです。是非チャレンジしましょう。」とステップアップを勧めることができると考える。

４．考察

今回、解析対象を 2014 年度から科学研究費が交付された基盤 A、 B、 C の代表者とした。何故直近の 2021 年に交付が開始される代表者を対象としなかったのか、また、2019 年度公募より基盤研究等における研究計画調書の「研究業績」欄が廃止された状況下で、研究業績（発表論文数）を最も重要視したクラス分類器の作成に意味があるのかと疑問を抱かれる方も多いと思う。一言述べておく必要があろう。本稿は下記の一連の調査研究の一つとして実施した。

① 申請前の公知のデータから、研究者に適した基盤 A、 B、 C の種目を提示可能なクラス分類器の作成。（本稿）
② 研究種目のステップアップをできる研究者を選択する方法の検討を行う。代表者のステップアップを 5、6 年にわたって追跡する必要がある。（一部本稿）
③ 研究費交付が発表論文数の増加に与える効果の調査。研究費交付前後の論文数を知る必要がある。（次報）

②と③に答えるためには 5、6 年の追跡期間が必要と考え、2014 年度開始の基盤研究を分析対象とした。

本項で開発した分類器は、既に基盤 A、 B、 C に採択された代表者が、どの種目に採択されているのかを予想する分類器になっている。本来であれば、不採択課題と採択課題のデータを用いて分類器の訓練を行うべきであるが、残念ながら不採択課題の情報が公開されていない。本稿では次善の策として、基盤 C の代表者が基盤 A に申請してもまず採択されることはなく、また、基盤 B の代表者が基盤 A に申請してもその採択率はさほど高くはないと考え、本報告の方法で分類器の訓練を行った。考え方としては、基盤 B、C を基盤 A の不採択課題として分類器の訓練を行っている。

平成 31（2019）年度の科学研究費の応募・審査時から「研究業績」欄が廃止された。このことは、日本学術振興会が“「研究業績を書かなくてよくなった」など、 誤った認識として捉えられている事例もあり”（日本学術振興会, 2019）と特別に書かなくてはならないほどの強いメッセージである。今後、わが国の論文発表数にどの様な影響があったか注意深く見守る必要がある。科学研究費の採否の観点からは、試作した分類器を 2020 年以降の採択課題に適用した時に、分類器の能力の低下が著しい時は、科学研究費の採択に論文が重要視されていないことになり、この意味からも興味深い。

ステップアップ可能な研究者を見出し、適切に助言してステップアップを実現させることは URA のやりがいではなかろうか。結果の 3.2.4 に書いたように、研究費獲得の確実性を望むためか、研究者は驚くほど保守的であり、ステップアップを試みる方が少ない。また、2014 年度採択の基盤 A の理系代表者 504 名の内 385 名、実に 76%の代表者が第３期中期目標に示された旧帝大を中心とした国③と理化学研究所や国立研究開発法人物質・材料研究機構などの有力国立研究所や研究機構の在籍者である。一方、基盤 B には国③以外の大学の研究者が多数おられる。ステップアップの観点から、今一度表２の AB 分類の混合行列を見ていただきたい。今回試作した分類器により、基盤 B の代表者の 33%は基盤 A に誤分類されている。この誤分類の結果が示すように基盤 B の代表者と基盤 A の代表者にかなりの重なりがあると考える。国③以外の大学の研究者で、基盤 A で

の採択の可能性がありながら基盤 B に申請されている方が多数おられるのではないだろうか。この様な研究者は URA の支援でステップアップが可能であると考える。

今回の報告した試みの大きな欠点は不採択課題の情報を用いていないことと、申請内容に関する情報が含まれていないことである。前者に関しては大学内に不採択課題の情報も蓄積されているので、その情報を用いればより良い教師データが作成できると考える。また、後者に関しては申請内容の優劣は申請者の研究実績に密接に関係するので、特徴量として論文数だけでなくその質も加味した h-index（Müller and Guido, 2017）等の指標を用いるのも一法かと考えて、現在データ収集を開始したとこである。二つの試みを融合させるとより優れた分類器を作成できると考える。不採択課題の情報にアクセスできる URA の方と是非共同研究させていただければと考えている。

最後に、分類器のプログラムコードとデータ分析法の詳細を知りたい方は、ご連絡いただければ提供いたします。ご連絡をお待ちしています。

引用文献・参考文献

久保琢也，伊藤広幸（2020）「基盤(A)にステップアップした研究者の研究費採択履歴の特徴」情報誌「大学評価と IR」第 11 号，15-44.

日本学術振興会（2019）「令和元（2019）年 9 月、“研究計画調書の変更（研究業績欄）について①、②、③」,『科研費の最近の動向及び令和 2 年度公募について』. https://www.jsps.go.jp/j-grantsinaid/06_jsps_info/g_190902_1/data/siryou2.pdf

文部科学省（2015）「28 年度概算要求：高等教育局主要事項」. https://www.mext.go.jp/component/b_menu/other/__icsFiles/afieldfile/2015/08/27/1361291_1.pdf

Hirsch, J. E. (2005) . An index to quantify an individual's scientific research output”, PNAS November 15, 2005 102 (46) 16569-16572.

Müller, A. C. and Guido, S. (2017). Introduction to Machine Learning with Python, O’Reilly Media.（邦訳）中田秀基『Python ではじめる機械学習』, オライリー・ジャパン.［サンプルコード URL］ https://github.com/amueller/introduction_to_ml_with_python

［受付：令和 3 年 7 月 31 日　受理：令和 3 年 9 月 1 日］