EMIRにおける中途退学の予測可能性

高松邦彦（神戸常盤大学）、村上勝彦（東京工科大学）、鷹尾和敬（神戸常盤大学）
旭潤一郎（神戸常盤大学）、桐村豪文（神戸常盤大学）、伴仲謙欣（神戸常盤大学）
野田育宏（神戸常盤大学）、光成研一郎（神戸常盤大学、京都大学）
中村忠司（神戸常盤大学）、中田康夫（神戸常盤大学）

要旨

神戸常盤大学では、平成 28 年度に IR（Institutional Research）推進室を設置した。そこでは、経営戦略のための IR だけではなく、エンロール・マネジメント（EM）を目的として、学内のさまざまな学生に関する情報を収集、整理、管理、提案を行っている。昨年度は、設置した年度であり、まず学内に散在している情報を把握するところから着手し、収集と整理を行った。

本年度、新たに IR 推進ユニットが設置された。IR 推進室が職員だけの部門であるのに対し、IR 推進ユニットは、教員と職員のメンバーから構成される教職協働のユニットである。IR 推進ユニットでは、本年度の課題として「学生の中途退学」に焦点を当て、解析を進めることになった。

現在、数値データで表現された項目が 1,246 項目、データ数（人数）が 2,163 件、データベースに存在する。我々は本研究のリサーチクエスチョンを「EMIR に機械学習を用いて中途退学予測の可能性について検討すること」とし、上記のデータを用いてエビデンス・ベースドに研究を行った。3 手法による機械学習を行った結果、ロジスティック回帰ではテスト正解率が 60%を超え、ランダムフォレストではテストの正解率が約 90%となった。

１．背景

2008（平成 20）年の中央審議会答申「学士課程教育の構築に向けて」において、学位授与、教育課程・編成の実施、入学者受け入れの 3 つの方針を主体として、教学経営の明確化や教職員の職能開発の確立等が示された [1)]。この答申以降、統計データなどの科学的根拠に基づいて判断を行う、いわゆる「エビデンス・ベースド（evidence based）」な考え方が大学にも求められるようになってきた。

この流れに沿って、2011（平成 23）年に学校教育法施行規則が改正され、各大学が公表すべき教育情報が明確化された。2012（平成 24）年の中央教育審議会答申「新たな未来を築くための大学教育の質的転換に向けて～生涯学び続け、主体的に考える力を育成する大学へ～」の中で、用語説明の「大学ポートレート」において、「大学ポートレート（仮称）の整備により、①大学が教育情報を用いて自らの活動状況を把握・分析し、改革につなげる（いわゆる IR（Institutional Research）機能の向上）」という IR の記述がみられ [2)]、これ以降、日本の大学における IR への関心が本格的に高まってきた。大学ポートレートについては、その後、2014（平成 26）年 7 月に、独立行政法人大学評価・学位授与機構に大学ポ

ートレート運営会議および同センターが設置され、私学については 2014（平成 26）年 10 月 6 日[3]、国公立については 2015（平成 27）年 3 月 10 日に大学ポートレート[4]が公開された。

この流れを汲み、神戸常盤大学では、まず準備段階として 2015（平成 27）年度に IR 委員会を設置し、次年度の 2016（平成 28）年度には IR 推進室を設置した。本学の IR は、経営戦略のための IR だけではなく、エンロール・マネジメント（EM）を目的として、学内のさまざまな学生に関する情報を収集、整理、管理、提案を行っている。昨年度は、設置された年度であり、まず学内に散在している情報を把握するところから着手し、収集と整理を行った。

本年度、新たに IR 推進ユニットを設置した。IR 推進室が職員だけの部門であるのに対し、IR 推進ユニットは、教員と職員のメンバーから構成される教職協働のユニットである。IR 推進ユニットでは、近年わが国で増加している「学生の中途退学」[5]を本年度の課題に据え、解析を進めることにした。

中途退学者については、これまでも学校基本調査報告のデータをもとにマクロなレベルで研究されている[5][6]。ミクロなレベルでは、退学防止を目的として学生相談的アプローチについて研究されている[7]。また、EMIR に関しては、山形大学[8]や京都光華女子大学[9]においても積極的に中途退学の問題について研究されている。科学研究費の研究においても、EMIR におけるデータに注目して中途退学の防止策が研究されている[10]。

「中途退学」は、「海外留学」や「他大学への編入」などの積極的・自発的な理由だけでなく、「経済的困難」や「就学意欲の低下」などの消極的・非自発的な理由とも関連している[5]。このように、中途退学にはさまざまな理由が存在するため、中途退学を予測することはこれまで非常に難しいとされてきた。この問題に対して、我々は本研究のリサーチクエスチョンを「EMIR に機械学習を用いて中途退学予測の可能性について検討すること」とし、エビデンス・ベースドに研究を行った。

２．方法

1）データの準備

本学の EMIR のデータについては、原則非公開となっている。そのため、本研究においては、IR 推進室より学籍番号、氏名などをすべて削除した匿名な状態で、また各項目についてもすべて番号へ変換して項目内容がわからないように秘匿した状態でデータを入手した。そのため、各項目が何を表しているのか、解析者には理解できないようになっている。ただし、機械学習の正解（中途退学者を示すデータ）が第 1 項目であること、また中途退学を直接表すデータは除外されていることのみ、IR 推進室から伝えられた。

データは、IR 推進室から csv 形式で入手した。入手したデータをタブ形式にして保存し、改行コードを Unix 形式に変換した。この際、改行が 2 回含まれているセルが複数あったため、それらを修正した。その後、漢字コードを nkf を用いて utf-8 に変換した。また、欠損データは 0 とした。

2）機械学習

解析は、mac OS X 10.11.6 で行った。解析には、Python（3.6.0）と Perl（5.18.2）を用いた。Python のライブラリとして numPy[11]、matplotlib[12]、scikit-learn[13]、pandas[14]を用いた。機械学習においては、トレーニングデータを 70%用い、テストデータを 30%用いた。

３．結果と考察

IR 推進室から入手したデータ数は、全部で 2,163 人分であった。各データに対して、1,246 項目が存在した。このうち、機械学習の正解（退学者を示すデータ）は第 1 項目であった。

我々のリサーチクエスチョン「EMIR に機械学習を用いて中途退学予測の可能性について検討すること」は、「2,163 人、1,246 項目のデータを用いて、中途退学するか、卒業するのかという 2 値判別について機械学習を用いて予測可能か？」と言い換えることができる。

機械学習の分野に no free lunch という定理がある [15]。この定理が示していることは、機械学習で注意すべき事象であり、それは「どのようなデータにおいても、高い精度を出せる万能な機械学習手法が存在しない」ということである。そこで我々は、3 つの手法を用いて機械学習を行った。第 1 と第 2 の手法は、ロジスティック回帰（logistic regression）を用いた機械学習である [16]。回帰分析は、目的変数（従属変数とも呼ばれ、本研究の場合は中途退学するか、卒業するかの 2 値）を、説明変数（独立変数とも呼ばれ、本研究の場合は 1,246 の項目）の式で当てはめ、予測、変数の効果の調査に用いられる手法である。ロジスティック回帰は、質的変数を線形に回帰させるためのアルゴリズムである。本研究の場合、2 値（離散的）で表された変数を、ロジスティック回帰を用いて線形に回帰させた。通常の回帰では、トレーニングデータに最も適合した結果を返すため、過度な学習（過学習）を引き起こす可能性がある。これを回避するため、正則化項（罰則項とも呼ばれる）として、L1 正則化項と L2 正則化項と呼ばれる方法を用いた。L1 正則化項は、係数の絶対値の和として、L2 正則化項は係数の 2 乗和を用いる。第 1 の手法は正則化として L2 を用い、第 2 の手法は正則化として L1 を用いた。ロジスティック回帰は、3 次式以上で判別させると、線形分離不可能な問題も解けるという特徴がある。

第 3 の手法は、ランダムフォレスト（random forest）である [17]。ランダムフォレストは、アンサンブル学習による機械学習アルゴリズムの 1 つである。複数の決定木（tree）を弱識別器として用い、その結果を統合（forest）して正しい結果を得るアルゴリズムである。パターン識別をはじめとして、回帰、クラスタリングに利用できる特徴をもっている。

表 1 に、3 手法の機械学習の結果を示す。第 1 の機械学習法では、トレーニングの正解率は 0.643 で、テストの正解率は 0.649 であった。トレーニングとテストの正解率に、ほとんど違いはみられなかった。第 2 の機械学習法では、トレーニングの正解率は 0.573 で、テストの正解率は 0.603 であった。トレーニングの正解率よりも、テストの正解率のほうが高くなった。第 1 と第 2 の手法の違いは、正則化の違いであった。この結果から、中途

退学については、L1 の正則化よりも L2 の正則化のほうが、正解率が高いということが明らかとなった。

第 3 の機械学習法では、トレーニングの正解率が 0.914 で、テストの正解率が 0.895 となった。ランダムフォレストの結果は、ロジスティック回帰よりも約 25 ポイントも正解率が高いことが明らかとなった。

表 1　3 手法の機械学習によるトレーニングとテストの正解率

正解率＼手法	1	2	3
トレーニングの正解率	0.643	0.573	0.914
テストの正解率	0.649	0.603	0.895

次に、3 手法において、予測に寄与している項目番号の上位を抜粋した（**表 2**）。第 1 と第 2 の手法については、各項目の係数を出力している。また、第 3 の手法については、特徴量の重要度を出力している。第 3 の手法での変数評価は、その情報がなくなると精度がどれだけ下がるかを規格化した数値であるため、数値が大きいなら良い（卒業できる）などの向きは存在せず、数値の大小は判断への影響の大きさを示している。

表 2　3 手法の機械学習による中途退学の予測に寄与している項目番号（上位のみを抜粋）

項目番号＼手法	1	2	3
933	0.117	0.021	
16	0.099	0.407	
351	0.083	0.039	
996	0.072	0.059	
1173	0.062	0.043	
590	0.061	0.077	
100	0.053	0.051	0.022
648	0.05	0.015	
594	0.049	0.047	
1219	0.048	0.056	
591	0.042	0.092	
355	0.04	0.023	
110	0.038	0.02	
1127	0.033	0.038	
107	0.03	0.021	
108	0.03	0.032	
337	0.029	0.023	
584	0.029	0.067	
202	0.026	0.024	
392	0.025	0.012	
99	0.023	0.024	
1136	0.022	0.021	
19	0.022	0.186	

項目番号＼手法	1	2	3
588	0.019	0.073	
1085	0.016	0.036	
38	0.016	0.011	
561	0.016	0.075	
101	0.015	0.069	0.011
1106	0.015	0.039	
350	0.015	0.013	
605	0.015	0.048	
867	0.014	0.017	
40	0.013	0.038	
1223	0.011	0.035	
306	0.011	0.028	
21	0.01	0.005	
1220	0.009	0.041	
226	0.007	0.017	
902	0.007	0.005	
1097	0.006	0.021	
609	0.005	0.032	
868	0.005	0.018	
926	0.004	0.013	
596	0.003	0.061	
80	0.003	-0.005	0.025
828	0.003	0.003	

項目番号＼手法	1	2	3
458	0.002	0.012	
338	0.001	0.017	
95	-0.002	-0.007	
27	-0.004	-0.015	
624	-0.004	-0.02	
546	-0.006	-0.007	
567	-0.01	-0.02	
279	-0.011	-0.012	
246	-0.013	-0.013	
1100	-0.016	-0.022	
39	-0.026	-0.067	
201	-0.027	-0.027	
1089	-0.031	-0.075	
22	-0.032	-0.038	
627	-0.035	-0.02	
586	-0.037	-0.021	
97	-0.042	-0.043	
84	-0.059	-0.036	
1083	-0.06	-0.051	
35	-0.11	-0.052	
83	-0.121	-0.2	

表 2 の作成は以下の手順で行った。まず、第 1 の手法では、係数が 0 でない項目が 67 項目存在したため、この 67 項目の係数を大きい順に並べた。この 67 項目全てに対し第 2 の手法の係数を抜粋して併記した。最後に、第 3 の手法については、第 1 の手法の 67 項目と重なる項目の特徴量を抜粋して併記した。

その結果、項目番号 80、100、101 が結果に影響していることが明らかになった。本研究より、機械学習によってある程度の精度をもって、中途退学の予測ができる可能性を見出すことができた。

今回の解析においては、卒業生と中途退学者のデータを用いて機械学習を行った。しかし、我々が最も興味があることは、中途退学者の予防である。そのためには、在学生のデータを機械学習に取り入れる必要がある。また、実際に EMIR において、機械学習を用いて中途退学の予測を行う場合には、現時点で分かってない項目データを精査する必要がある。項目データには、中途退学を直接表すデータが含まれていないことは明らかになっている。しかし、項目群の中に、中途退学に間接的に依存したデータが含まれている可能性は否定できない。運用する場合には、IR 推進室の担当者が、解析対象となるデータを精査する必要があるだろう。近い将来、卒業生、在学生、そして中途退学者のデータを取り入れて機械学習を行い、在学生の中途退学者を予測できる可能性についてさらに研究を推進していく予定である。

【参考文献】

1) 中央審議会. “学士課程教育の構築に向けて”. http://www.mext.go.jp/b_menu/shingi/chukyo/chukyo4/houkoku/080410.htm, (参照 2017-06-13).
2) 中央教育審議会. “新たな未来を築くための大学教育の質的転換に向けて～生涯学び続け、主体的に考える力を育成する大学へ～”. http://www.mext.go.jp/component/b_menu/shingi/toushin/__icsFiles/afieldfile/2012/10/04/1325048_3.pdf, (参照 2017-06-13).
3) 独立行政法人大学改革支援・学位授与機構大学ポートレートセンター. “大学ポートレート（私学版）”. http://up-j.shigaku.go.jp/, (参照 2017-06-13).
4) 独立行政法人大学改革支援・学位授与機構大学ポートレートセンター. “大学ポートレート”. http://portraits.niad.ac.jp/, (参照 2017-06-13).
5) 姉川恭子. 大学の学習・生活環境と退学率の要因分析. 経済論究. 2014, vol. 149, p. 1–16.
6) 丸山文裕. 大学退学に対する大学環境要因の影響力の分析. 教育社会学研究. 1984, vol. 39, p. 140–153.
7) 窪内, 節子. 大学退学とその防止に繋がるこれからの新入生への学生相談的アプローチのあり方. 山梨英和大学紀要. 2009, vol. 8, p. 9–17. http://ci.nii.ac.jp/naid/110007616547/ja/, (参照 2017-06-13).
8) 福島真司. 「総合的学生情報データ分析システム」の構築 山形大学におけるエンロールメント・マネジメントとインスティテューショナル・リサーチ. 情報管理. 2015, vol.

58, p. 2–11.

9) 山本嘉一郎. エンロールメント・マネジメントを効果的に進めるためのIR について. 京都光華女子大学研究紀要. 2013, vol. 51, p. 89–98. http://ci.nii.ac.jp/naid/110009684596/ja/, (参照 2017-06-13).
10) 橋本智也. “データに基づく大学生の中途退学防止策(IR)のモデル構築：日米の制度差に着目して”. https://kaken.nii.ac.jp/ja/report/KAKENHI-PROJECT-15H00090/15H000902015jisseki/, (参照 2017-06-13).
11) Van Der Walt, Stéfan, Colbert, S.Chris, Varoquaux, Gaël. The NumPy array: A structure for efficient numerical computation. Computing in Science and Engineering. 2011, vol. 13, no. 2, p. 22–30.
12) Hunter, John D. Matplotlib: A 2D graphics environment. Computing in Science and Engineering. 2007, vol. 9, no. 3, p. 99–104.
13) Pedregosa, Fabian, Varoquaux, G. Scikit-learn: Machine learning in Python. 2011, 2825-2830p., ISBN9781783281930. http://dl.acm.org/citation.cfm?id=2078195.
14) McKinney, Wes. Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference. 2010, vol. 1697900, no. Scipy, p. 51–56. http://conference.scipy.org/proceedings/scipy2010/mckinney.html.
15) Wolpert, Dh. No free lunch theorems for search. Most. 1995, p. 1–38. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.47.7505&rep=rep1&type=pdf.
16) Cox, D. .. The Regression Analysis of Binary Sequences. Journal of the Royal Statistical Society. 1958, vol. 20, no. 2, p. 215–242.
17) Breiman, L. Random Forests. Machine Learning. 2001, vol. 45, no. 1, p. 5–32.