機械学習を用いた学力進捗予測の可能性

高松邦彦（神戸常盤大学）、村上勝彦（東京大学）、鷹尾和敬（神戸常盤大学）
村瀬有紀（神戸常盤大学）、深川大（神戸常盤大学）、旭潤一郎（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）
大森雅人（神戸常盤大学）、中田康夫（神戸常盤大学）

1. 要旨

神戸常盤大学では、2016(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。

2016(平成 28)年度は、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。2017(平成 29)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「学力進捗」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学力進捗予測の可能性について検討すること」とし、上記の一部のデータを用いてエビデンス・ベースドに研究を行ったので紹介する。

2. 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。2017（平成 29）年度には、IR 推進ユニットを設置した。IR 推進ユニットは、IR 推進室が職員だけの部門であるのに対し、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。

昨年度 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」[2]を本年度の課題に据え解析を進め、いくつか報告してきた[3], [4], [5]。「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[2]。

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。我々は、その研究において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」とした。その結論、中途退学はある程度の予測精度で予測可能であり、また顕著な理由（項目）は発見できないことを報告した[3], [4], [5]。また、機械学習を EMIR に適用し、中途退学の予測を行うという研究は、我々以外にも行われており機械学習が有効な解析手段であることは明らかになりつつある[6], [7], [8]。

近年、東京理科大学の調査によると、卒業時の成績は、1 年時の成績と相関があることが報告されている。毎日新聞の記事では、「大学卒業時の成績は 1 年終了時の成績とほぼ一致し、入学試験の結果とは相関関係がみられないことが、東京理科大学（東京都新宿区）が同大の学生を対象に実施した調査で明らかになった。担当した山本誠副学長は「『特に 1 年の 6 月第 1 週の出欠状況が、その後の学生生活を左右する』と話している」と紹介されている[9]。

本学の IR 推進室の解析でも、類似な傾向が見られた（data not shown）。そこで我々は、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学力進捗予測の可能性について検討すること」として、エビデンス・ベースドの研究を行ったので報告する。具体的には、「1 年時科目別成績データ（1,155 人、118 科目）を用いて、国家試験に合格するかどうか予測可能か？」について機械学習を用いて解析した。

3. 方法

1）データの準備

本学の EMIR のデータについては、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データを入手した。

データは、IR 推進室からエクセル形式で入手した。データに含まれる学生数（ケース数）は 1,155 人、説明変数（1 年時の科目の成績等）は 118 変数であった。科目ごとの数値範囲は基本的に点数であり、0～100 の範囲である。科目登録していない学生は 0 点としている。よって欠損データは無い。これが学生数としては最も多い。0 より大きい数値となっている学生の場合、大多数は 60 から 100 の範囲で、60 以下は極端に少ない。登録した学生が（離脱をのぞいて）0 点をつけられることは基本的に無いが、その事象が起こったと

しても解析に対する支障は特にない。

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）を用いた。Python のライブラリとして numPy[10]、matplotlib[11]、scikit-learn[12]、pandas[13]を使用した。全ケースのうち約 70%を学習データ（具体的な予測モデルの構築）として用い、残り約 30％をテストデータ（学習に使っていない未知のデータであり、予測モデルの評価に利用）として用いた。割当については、変数値の偏りがなるべくないように scikit-learn による仕組みで割り当てている。

4. 結果と考察

IR 推進室から入手したデータ数は、全部で 1,155 人分であった。本学は、短期大学部に口腔保健学科が、大学の保健科学部に医療検査学科と看護学科が、大学の教育学部にこども教育学科が設置されている。

今回の解析に使用したデータは、全 4 学科のデータとなっている。この全 4 学科の各データに対して、1 年時の成績の項目（科目）が 118 項目存在した。この 118 項目（科目）は、全学科の科目を合計し、重複を抜いたものと同等である。

我々のリサーチクエスチョン「EMIR に機械学習を用いて学力進捗予測の可能性について検討すること」であり、より具体的には「1 年時科目別成績データ（1,155 人、118 科目）を用いて、国家試験に合格するかどうかが予測可能か？」であった。そこで、118 個の説明変数 X を用いて予測対象を目的変数 Y（0 または 1）とした合格・不合格のカテゴリ予測問題を設定した。

国家試験の合格・不合格の予測方法については、多くのタイプのデータで比較的安定的に高精度の予測ができるといわれる Random Forest 法による予測をおこなった[14]。この方法は簡単にいえば、「学習データの一部から決定木を作成すること」を繰り返すことによって、互いに相関の弱い決定木を数多く作成し、それらの多数決で最終予測する方法である。

表 1 に、機械学習の結果を示す。第 1 の機械学習法では、学習の正解率は 0.966 で、テストの正解率は 0.913 であった。学習とテストの正解率はどちらも高く、過学習の可能性はわずかであると見られる。

表 1　機械学習（Random Forest 法）の正解率

評価に使用したデータ	正解率
学習データ	0.966
テストデータ	0.913

次に、テストデータについて、合格と予測した数を分母とした正解率（精度, Precision, P）と、実際に合格した数を分母とした正解率（再現率, Recall, R）を比較した（表 2）。そしてこれらの調和平均である F 値を計算した（表２）。F 値の定義は 1/F = (1/P + 1/R) /2 である。今回のモデルでは合格と予測した場合には 96.7%正しいと期待できる。再現率をみると合格者の 81.7%を正しく予測できたことからみても、非常に高い正解率である。

表 2　2 種類の正解率とその調和平均

Precision（精度）	Recall（再現率）	F 値（調和平均）
0.967	0.817	0.885

はじめに提示したリサーチクエスチョン「1 年時科目別成績データ（1,155 人、118 科目）を用いて、国家試験に合格するかどうかが予測可能か？」については、予測できる可能性が高いといえる結果であろう。予測モデルに利用したデータは 1 年生時点の各科目の成績であるから、1 年生の時点という早い時期に正確な判断基準をもとにした指導が可能になるのではないか、と期待できる。

【参考文献】

[1] 中央審議会, “学士課程教育の構築に向けて,” 2008. [Online]. Available: http://www.mext.go.jp/b_menu/shingi/chukyo/chukyo4/houkoku/080410.htm. [Accessed: 13-Jun-2017].

[2] 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, vol. 149, pp. 1–16, 2014.

[3] 高松邦彦, 村上勝彦, 鷹尾和敬, 旭潤一郎, 桐村豪文, 伴仲謙欣, 野田育宏, 光成研一郎, 中村忠司, and 中田康夫, “機械学習による中途退学の予測可能性,” *第6回 大学情報・期間調査研究会*, pp. 60–65, 2017.

[4] 高松邦彦, 村上勝彦, 鷹尾和敬, 旭潤一郎, 桐村豪文, 伴仲謙欣, 野田育宏, 光成研一郎, 中村忠司, and 中田康夫, “機械学習による中途退学の予測可能性,” *計測自動制御学会 システム・情報部門, 2017*, pp. 745–746, 2017.

[5] K. Murakami, K. Takamatsu, Y. Kozaki, A. Kishida, K. Bannaka, I. Noda, J. Asahi, K. Takao, K. Mitsunari, T. Nakamura, and Y. Nakata, “Predicting the Probability of Student Dropout through EMIR Using Data from Current and Graduate Students,” *Adv. Appl. Informatics (IIAI-AAI), 2018 7th IIAI Int. Congr. on. IEEE*, p. in press, 2018.

[6] N. Kondo, M. Okubo, and T. Hatanaka, “Early Detection of At-Risk Students Using Machine Learning Based on LMS Log Data,” in *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, 2017.

[7] 近藤伸彦 and 松田岳士, “教学IRにおける予測モデル活用の枠組み,” *第6回 大学情報・期間調査研究会*, pp. 42–47, 2017.

[8] 近藤伸彦 and 畠中利治, “教学 IR における LMS ログデータ活用の試み,” *計測自動制御学会 システム・情報部門, 2017*, p. 752, 2017.

[9] 毎日新聞, “大学成績 1年で決まる？ 卒業時と一致 東京理科大調査,” 毎日新聞, 2016. [Online]. Available: https://mainichi.jp/articles/20160603/k00/00m/040/141000c. [Accessed: 01-Jun-2018].

[10] S. Van Der Walt, S. C. Colbert, and G. Varoquaux, “The NumPy array: A structure for efficient numerical computation,” *Comput. Sci. Eng.*, vol. 13, no. 2, pp. 22–30, 2011.

[11] J. D. Hunter, “Matplotlib: A 2D graphics environment,” *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 99–104, 2007.

[12] F. Pedregosa and G. Varoquaux, *Scikit-learn: Machine learning in Python*, vol. 12. 2011.

[13] W. McKinney, “Data Structures for Statistical Computing in Python,” *Proc. 9th Python Sci. Conf.*, vol. 1697900, no. Scipy, pp. 51–56, 2010.

[14] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 2, pp. 5–32, 2001.