教学 IR における予測モデル活用の枠組み

近藤伸彦，松田岳士（首都大学東京）

１．はじめに

近年，「教育ビッグデータ」「学習ビッグデータ」という語が示すように，教育機関において多種多様なデータが大規模に蓄積される環境が整いつつあり，データを活用した教育改善がさまざまに試みられている．学習者に関するデータの分析をもとに，学習者や学習環境の理解と適応的な学習支援や介入を行うことに関する研究分野として，この数年の間にラーニングアナリティクスおよびエデュケーショナル・データマイニングが発展している(1)ことなどがその一例である．その一方，日本において組織的導入と発展が続く IR（Institutional Research）においても，その対象とするデータは情報技術の発展にともなって増大し続けている．とくに，教育に関連する教学 IR においては，こうしたデータの有効活用にむけて，ラーニングアナリティクス等の教育ビッグデータ分析の関連分野との有機的な統合が望まれている(2)．

ラーニングアナリティクス等では，学習に関するなんらかの「予測」が主要な役割を果たすことが多く，そのための予測モデルは一般に機械学習やデータマイニングの手法を用いて構築される．本稿では，教学 IR における予測モデルの活用に焦点をあてる．これまでにも，教学 IR における予測モデルの活用はさまざまに試みられているが，使用可能なデータの種類や粒度などは機関ごとに大きく異なるため，ある機関における研究結果が他の機関に直接的に応用可能であることは必ずしも多くないと考えられる．本稿では，教学 IR において予測モデルを活用するためには，個々の機関が自身の文脈に応じて予測モデルの構築と評価を行うことが必要であるという立場をとり，そのための枠組みについてまとめる．

２．データにもとづく教育改善と予測モデル

データにもとづく教育改善における「予測」の果たす役割は大きくなりつつあり，その実践事例も多い．Brooks and Thompson は，ラーニングアナリティクスやエデュケーショナル・データマイニングにおける予測モデルの活用について，その構築プロセスや具体的なモデルの種類，実践例，課題等についてまとめている(3)．予測対象には，学生の成功（academic success）や学習成果，教育方法による教育成果，リテンションや学習上のリスク（academic risk）などが一般に想定されるとあり，予測モデル（predictive model）と説明モデル（explanatory model）を区別する必要性についても言及されている．予測モデルの目的が未知のデータに対する予測を与えることであるのに対し，説明モデルはこれまでの結果や現象に対する原因を分析することがその目的である．

教学 IR においては，従来は説明モデルによる分析や可視化にもとづく意思決定支援が主要な機能であったと考えられるが，近年では教学データにもとづくなんらかの予測に関する研究が多く行われていることから，予測モデルにもとづく種々の予測が教学 IR においてもその重要性を増していると考えられる．教学 IR における予測の試みには，たとえ

ば単位修得状況の予測[4]や，留年・退学の予測についての数多くの報告がある[5][6][7][8][9]．

これらの事例は，特定の機関（大学）におけるデータを用いたものであり，それぞれ使用されているデータや予測モデルなどは異なる．これは，教学 IR において使用可能なデータの種類や蓄積の頻度などが機関ごとに異なることから当然に生じることである．また，機関の種別（国公私立など）や規模，構成される学問分野，3 つのポリシーなど，機関の状況は千差万別であり，ある機関における研究結果が他の機関に直接的に応用できるとは考えにくい．そのため，予測モデルの活用を実質化するためには，各機関が自身の文脈に応じて予測モデルを構築し，その結果を評価することが必要になると考えられる．

３．教学 IR における予測モデル活用の枠組み

本章では，教学 IR 担当者が自ら予測モデルを活用することを想定して，予測モデル活用のための枠組みについて整理する．

３．１ 予測モデルの構築

予測モデルは，予測対象が連続値であれば回帰（regression）モデル，離散値であれば分類（classification）モデルが用いられるなどの違いはあるが，基本的にはモデルへの入力である説明変数から出力である目的変数への写像を数学的に表したものである．予測モデルは，あらかじめ用意されたデータを用いて，データがよりよく予測できることを是とする評価基準にもとづいて統計的に学習（training）される．

予測モデルを用いてなんらかの予測を行う場合，モデルの学習に用いるデータはすでにあるものとして，少なくとも次の 5 点を定める必要がある．

1) 目的変数（出力）にどの変数を用いるか．
2) 説明変数（入力）にどの変数を用いるか．
3) どの予測モデルを用いるか．
4) 予測モデルの構造やパラメータ，学習アルゴリズムをどのように定めるか．
5) 予測モデルの性能をどのように評価するか．

次節以降で，これらの点を定めるうえでのポイントをまとめる．

３．２ 予測モデルに用いる変数と学修ライフログ

3.1 節の 1) および 2) を適切に定めるためには，膨大な教学データをあらかじめ整理しておく必要がある．学内にどのようなデータがあり，個々のデータがどのような関係にあるのかを明らかにして，操作可能にしておくことが必要である．

教学 IR において予測モデルを活用することを考える場合，その多くは学生に関するなんらかの予測が想定されると考えられる．近藤・畠中は，教学 IR において学生のデータを整理するための概念として，学内に散在する大規模なデータを一元集約し学生ごとに時系列に整理した「学修ライフログ」の概念を提示している[9]．これは，時間進行にしたがってライフログのように教学データが蓄積されていくさまを模式的に表わしたものである．学修ライフログのイメージを図 1 に示す．これは概念であるので，実際のデータベースの

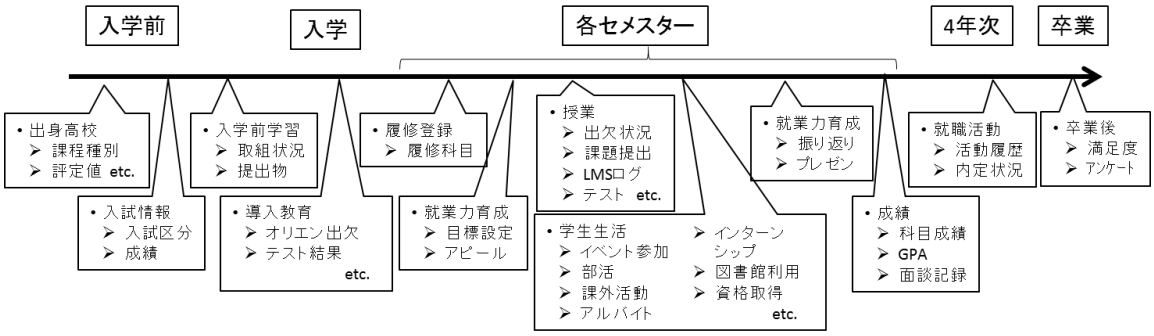

図１　学修ライフログによるデータの時系列整理（文献[9]より引用）

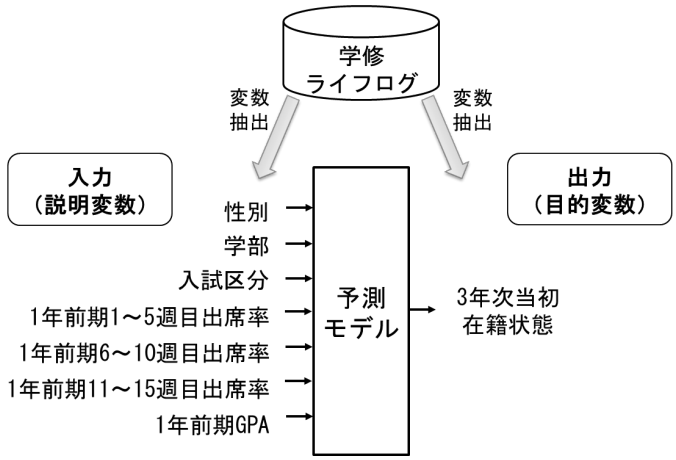

図２　学修ライフログからの予測モデル変数抽出

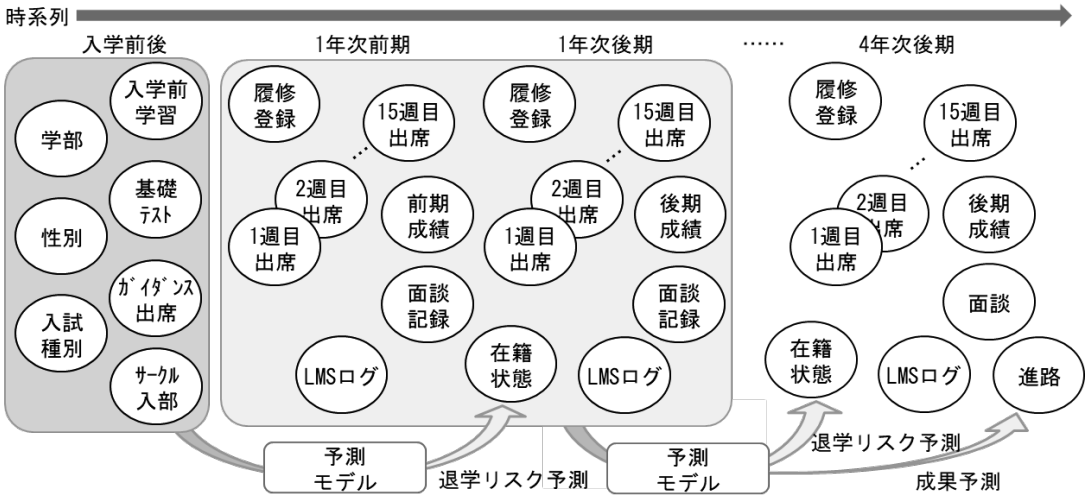

図３　時間経過にともなう予測モデル変数の任意抽出

仕様とは切り分けたメタデータとして適宜整理しておけばよい．このような概念にしたがって，時系列上の変数間の関係を整理しておくと，予測モデルのための変数抽出が容易になると考えられる．

何を予測したいかという目的を明確にしたのち，学修ライフログとして整理したデータベースから，まずは 1）の目的変数が抽出できる．2）の説明変数については，目的変数よりも時系列上過去にあたる変数が抽出可能であるが，「どの時点で予測を行いたいか」によって変数の候補は絞られる．たとえば，1 年次の春の大型連休までに，退学リスクのある学生を予測したいとすれば，それまでに入手可能な変数を抽出候補とすればよい．学修ライフログから変数抽出をするイメージを図 2 に示す．実際には，変数の離散化や正規化といったデータの前処理や，元のさまざまなデータを結合・変換してなんらかの特徴量を抽出することも必要になることがほとんどである．

「どの時点でどのような予測を行いたいか」という問題設定にはさまざまなものが考えられる．図 3 は，学修ライフログの時系列上で，それぞれの問題設定に応じて説明変数・目的変数を抽出しているようすを示したものである．

３．３ 変数の特徴にあわせた変数抽出

文献[10]では，予測モデルを構成する変数が 4 つに分類されている．以下にこの 4 分類を引用する．

【Type 1: 変更不可能な個人属性】性別，年齢，出身地等のデモグラフィック属性や，所属学部，入試種別など入学前に確定している変数．

【Type 2: 学習成果】GPA（Grade Point Average）やテストの点数，修得単位数，就職などの学習成果に関する変数．

【Type 3: 行動の結果や状態】出席率や課題提出率，サークル在籍の有無，休退学や留学等の在籍状態など，学生の行動の結果や状態を示す変数．

【Type 4: 大学からの介入の有無】アカデミックアドバイザーや学習支援員からの連絡など，大学からの修学上の介入の有無を示す変数．

Type 1 は入学以降変更不可能なもの，Type 2 は学生が目標設定できても完全に制御するのは困難なもの，Type 3 は学生のアクション次第で能動的に制御可能であるもの，Type 4 は学生側ではなく大学側からのアクションの有無を示すものである．

学修ライフログから変数を抽出する際，それがどの Type にあたるかを明確にしておけば，「予測」の結果によって機関がどのようなアクションを起こせるかの指針ともなると考えられる．文献[10]にあるように，たとえば修学支援対象者のリストアップを念頭に置いた場合，Type 2 や Type 3 の変数を目的変数とすれば，学習成果やドロップアウトリスクを予測できる．今後の学生の行動指針を与えたり特定の介入をすべきかどうかを判断したりする場合には，時系列のうえで未来に属する Type 3 や Type 4 の変数へ仮の値をセットして予測を行う．一方で，直接の修学支援ではなく，現状分析としての教学 IR に活用するならば，Type 1 の変数に着目して学生の属性についての分析を行ったり，Type 4 の変数に着目して学習支援施策の効果検証を行ったりすることも考えられる．これは予測モデルというより説明モデルとしての機能に近い．

３．４ 予測モデルの選択，学習，評価

3.1 節の 3)～5) は，教学データに限らず，予測モデルの学習（統計的学習）に共通の課題である．

3) に関しては，説明変数から目的変数への写像の表現方法の違いによってさまざまなモデルが存在している．先述の文献[3]においては，教育データの予測モデルによく使われるものとして，線形回帰，ロジスティック回帰，最近傍法，決定木，ナイーブベイズ分類器，ベイジアンネットワーク，サポートベクターマシン，ニューラルネットワーク，アンサンブル法が紹介されている．近年の機械学習の発展により，R や Python などのプログラミング言語において，信頼性が高くかつ豊富な機械学習パッケージが利用可能であり，これらを活用することで容易に多様な予測モデルの構築が可能である．対象とするデータの性質や本質的な構造によって，どのモデルが良い性能を示すかは異なるため，用いるモデルの選択には予備的な調査が必要となることが一般的である．

4) は，予測モデルの複雑さや内在するパラメータ，統計的学習のアルゴリズムに関するもので，モデルの汎化能力（学習に用いていない未知のデータを予測する能力）に影響を与える要素についてのものであり，一定の知識や経験，試行錯誤が要求されるものである．一般に，情報量規準や交差検証などの方法によって検討することが求められる．

5) は，予測モデルの性能をどのような指標をもって評価するかということであり，最終的にどのモデルを用いるかの判断につながるものである．上で述べた汎化能力を考慮した何らかの評価指標にもとづいてモデルの評価を行うことが一般的である．教育で用いられる予測モデルの多くは，目的変数が離散値をとる分類モデルであることが多く，パターン認識の分野でよく用いられる評価指標がしばしば用いられる．たとえば，適合率（precision; モデルが予測した分類ラベルが真の値と等しい割合），再現率（recall; 真の値のうち，予測モデルによって正しく分類された割合），F 値（precision と recall の調和平均），ROC 曲線（受信者動作特性曲線; 分類の閾値を変化させたときの偽陽性率と真陽性率を 2 次元でプロットしたもの）などがよく用いられる．これらの評価指標のもつ意味を理解したうえで，予測によって実現したい目的に応じて適切に評価指標を選択し，多面的に検討することが重要である．

４ 教学 IR における予測モデル構築の支援ツールと事例共有

第 2 章で述べたように，教学 IR において予測モデルを活用するためには，機関の文脈に応じて個々の機関が予測モデルの構築と評価を行うことが必要であると考えられる．教学 IR における予測モデル活用をより活性化させるためのひとつの方法として，必要以上に高度な専門的知識がなくても扱うことができかつ理論的に適切な方法をとれるような，予測モデル構築を支援する汎用的なツールを作成しこれを広く共有することと，これを用いた結果や成果を機関内外で手軽に共有するしくみを構築し実質化することが考えられる．簡易なツールにより容易になる実践事例の共有を通じて，教学 IR における予測モデル活用のあり方の議論が活発化するとともに教学 IR 担当者のもつ予測モデルについてのリテラシーが向上し，各機関がそれぞれ独自に，適切に予測モデルを活用できる状態になることが期待される．

予測モデル構築支援ツールの要件は，3.1 節に示した 5 点を支援する機能を有すること

であり，3.2 節以降に述べたような点をおさえたものであることが重要であると考えている．本発表においてはそのプロトタイプを紹介する予定である．

5 おわりに

本稿では，教学 IR における予測モデル活用のための枠組みについてまとめた．教学 IR における予測モデル活用を活性化させるには，機関の文脈に応じて個々の機関が予測モデルの構築と評価を行い，その知見を共有することが望ましいと考え，これを支援する汎用ツールの構築を構想している．まずは簡便なツールを広く使える状態とし，どのようなことができるかを手軽に試せる環境をつくり，それぞれの機関内での試行と検証，そして可能な範囲での事例共有を行い，必要十分な機能をもつツールへと発展させるとともに，予測モデル活用の要点とノウハウを多くの教学 IR 担当者が持つことができる状態をめざしたいと考えている．

謝辞

本研究の一部は JSPS 科研費 JP16K16331 および JP16H03082，ならびに首都大学東京傾斜的研究費の助成を受けた．

【参考文献】

(1) G. Siemens and R .S. J. d. Baker, “Learning Analytics and Educational Data Mining: Towards Communication and Collaboration”, Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, pp. 252-254, 2012.
(2) 船守美穂, “デジタル技術は高等教育のマス化問題を救えるか？―MOOCs, 教育のビッグデータ，教学 IR の模索”，情報知識学会誌，Vol.24，No.4，pp.424-436, 2014.
(3) C. Brooks and C. Thompson, “Predictive Modelling in Teaching and Learning”, Handbook of Learning Analytics, pp. 61-68, 2017.
(4) 雨森聡，松田岳士，森朋子，“教学 IR の一方略：島根大学の事例を用いて”，京都大学高等教育研究，第 18 号, pp.1-10, 2012.
(5) 大友愛子，岩山豊，毛利隆夫，“学内データの活用～大学における IR（Institutional Research）への取組み～”, FUJITSU, Vol.65, No.3, pp.41-47, 2014.
(6) 嶌田敏行，“留年してしまう学生の効率的・効果的な検出方法についての検討”，大学評価と IR，第 4 号, pp.18-25, 2015.
(7) 藤原宏司，“学業を中断する学生の予測モデル構築について”，大学評価と IR，第 5 号, pp.8-22, 2016.
(8) 竹橋洋毅，藤田敦，杉本雅彦，藤本昌樹，近藤俊明，“退学者予測における GPA と欠席率の貢献度”，大学評価と IR，第 5 号, pp.28-35, 2016.
(9) 近藤伸彦，畠中利治, “学士課程における大規模データに基づく学修状態のモデル化”，教育システム情報学会誌, Vol.33, No.2, pp.94-103, 2016.
(10) 近藤伸彦，畠中利治, “ベイジアンネットワークによる修学状態推移モデルの構築”，日本教育工学会論文誌，第 41 巻第 3 号, 2017（採録決定）．