

確率モデルを用いた学修状態のモデル化と学修支援

Enrollment Management Using Probabilistic Model Based on Student Data

近藤 伸彦^{*1}, 畠中 利治^{*2}

Nobuhiko KONDO^{*1}, Toshiharu HATANAKA^{*2}

^{*1} 大手前大学 CELL 教育研究所

^{*1} CELL Institute for Educational Development, Otemae University

^{*2} 大阪大学大学院情報科学研究科

^{*2} Graduate School of Information Science and Technology, Osaka University

Email: nkondo7@otemae.ac.jp

あらまし：ユニバーサル化した高等教育における質保証や教育改善を行ううえでは、学士課程における学生の学修状況を適切に把握し、個に応じた適時的な学修支援を行うことが必要である。本研究では、学士課程において蓄積される大規模データを活用して、学生の学士課程にわたる学修状態をベイジアンネットワークで確率モデル化することを考え、その学修支援への活用について検討する。

キーワード：Learning Analytics, 教学 IR, 学修データ, 確率モデル, ベイジアンネットワーク, 学修支援

1. はじめに

ユニバーサル化した高等教育においては、多様な学生それぞれの個に応じた学修支援を行うことがますます重要となっている。教学 IR(Institutional Research)が近年急速に注目を集めていることも、そのひとつの証左である。また、近年の「教育ビッグデータ」という語に象徴されるように、教育の世界において、大規模なデータを知的に活用して教育の改善をめざす分野がめざましく発展している。

本研究では、学士課程において蓄積される学修データをもとに、学修状態をベイジアンネットワークによって確率モデル化することを考え、これを学修支援へ活用することについて検討する。

2. 先行事例

この数年、大規模なデータを活用して教育の改善をめざす Learning Analytics(LA)や Educational Data Mining(EDM)とよばれる分野が急速な発展を遂げている⁽¹⁾。その中には、ドロップアウトする、あるいは逆に高い学習成果をあげる学生のパターンを機械学習により発見する研究が散見される。たとえば Jayaprakash らは、デモグラフィック属性と LMS のログデータを活用し、授業の離脱リスクのある学生 (At-risk students)を早期発見して適切な介入を行うシステムを構築している⁽²⁾。これらの例は、LA や EDM の有用性を示唆している。教学 IR の観点においても、こうした数理的手法による教育ビッグデータの活用はますます重要性を増していくものと思われる。

3. 確率モデルを用いた学修状態のモデル化

3.1 ベイジアンネットワークによるモデル化

ベイジアンネットワークは、確率変数をノードとする非循環有向グラフネットワーク構造と条件付確率パラメータ集合による離散確率分布の近似モデルである。有向リンクは確率変数間の依存関係を示し、

条件付確率表 (CPT) により定量的に表される⁽³⁾。

本研究では、ベイジアンネットワークを用いて学士課程における学修状態の時系列的な確率構造をモデル化することを提案する。学生に関する大規模データから抽出した適当な特徴量をノードとし、時系列的順序を考慮しながらこれを結合してネットワーク構造を仮定し、実際の学生データから条件付確率を推定することでモデルを構築する。学士課程の任意の時点において、それまでのデータをエビデンスとして与えることで、以降の学修状態を確率推論したり、それに基づいて退学リスクのある学生を早期発見したりするなど、さまざまな活用が考えられる。

3.2 数値実験 1：在学状況のパターン分類

数値実験では、X 大学（教養教育型の小規模私立大学）における 2009～2013 年度入学生のデータを用いた。着目した特徴量を表 1 に示す。モデルの学習には 2009～2012 年度入学生のデータ (N=3,092) を、テストには 2013 年度入学生のデータ (N=598) を使用した。数値実験には Weka⁽⁴⁾を用いて実行した。

時系列的な順序をもとに、構造の複雑さが異なる 2 種類のモデルを図 1 のように定め、それぞれ学習データから CPT を求めた。なおその際に、連続値の変数については、いくつかのクラスに離散化した。

表 1 使用した特徴量

特徴量 (確率変数)	値
X ₁ : 性別	名義尺度, 2 クラス
X ₂ : 学部	名義尺度, 3 クラス
X ₃ : 入試種別	名義尺度, 7 クラス
X ₄ : 入学前学習課題提出度	順序尺度, 2 クラス
X ₅ : 新入生オリエンテーション出席度	順序尺度, 3 クラス
X ₆ : 1 年春学期 1～5 週必修出席率	連続値, [0, 1]
X ₇ : 1 年春学期 6～10 週必修出席率	連続値, [0, 1]
X ₈ : 1 年春学期 11～15 週必修出席率	連続値, [0, 1]
X ₉ : 1 年次春学期 GPA	連続値, [0, 4]
X ₁₀ : 3 年次 4 月在籍状況	名義尺度, 2 クラス

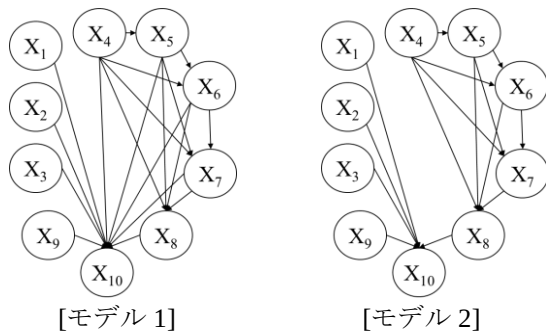


図 1 ベイジアンネットワークの構造

これら 2 種類のモデルを用いて、 X_{10} (3 年次 4 月在籍状況) を目的変数としてパターン分類を行った。 X_{10} は、在籍中である“OK”と、休学・退学・除籍となっている“not OK”の 2 クラスである。

データを正しく分類した割合 (正確度: Accuracy) は、学習データについてはモデル 1 が 95.41%, モデル 2 が 89.94% であるのに対し、テストデータについてはモデル 1 が 84.28%, モデル 2 が 85.62% であった。モデル 2 に比べて複雑な構造をもつモデル 1 は過学習していることがわかる。

学習データ、テストデータそれぞれに対する分類結果の混同行列を表 2~5 に示す。適合率(Precision)はモデルの予測が正解である割合、再現率(Recall)は真のクラスが正しく分類されている割合を表わす。ドロップアウトのリスクがある学生を早期発見するという観点からは、“not OK”クラスについての

表 2 学習データに対する結果 (モデル 1)

		予測		再現率 (Recall)
		OK	not OK	
実 際	OK	2,607	6	0.998
	not OK	136	343	0.716
適合率(Precision)		0.950	0.982	

表 3 テストデータに対する結果 (モデル 1)

		予測		再現率 (Recall)
		OK	not OK	
実 際	OK	491	8	0.984
	not OK	86	13	0.131
適合率(Precision)		0.851	0.619	

表 4 学習データに対する結果 (モデル 2)

		予測		再現率 (Recall)
		OK	not OK	
実 際	OK	2,529	84	0.968
	not OK	227	252	0.526
適合率(Precision)		0.918	0.750	

表 5 テストデータに対する結果 (モデル 2)

		予測		再現率 (Recall)
		OK	not OK	
実 際	OK	466	33	0.934
	not OK	53	46	0.465
適合率(Precision)		0.898	0.582	

Recall を高めることが望ましいと考えられる。この点においても、モデル 2 はモデル 1 に比べてテストデータにおいて高い Recall を示していることから、より汎化能力に優れていることがわかる。

3.3 数値実験 2: 時系列における段階的な確率推論

$X_1 \sim X_{10}$ は、学士課程が進行するにしたがって段階的に値が確定していくが、ある時点において、確定した変数のみを用いて確率推論することも可能である。数値例として、表 6 のように値をもつ 2013 年度入学のある学生について、時系列の順にこれらの値をエビデンスとしてモデルに与えるときに、確率推論される X_{10} の CPT がどのように変化するかを表 7 に示す。ここでは「モデル 2」を用いている。データが確定するにしたがって CPT が変化し、“not OK”の確率が高まっていくようすがわかる。同様に、他の変数の CPT に着目することも可能である。

表 6 2013 年度のある学生のデータ

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
b	a	d	low	low	0.18	0.0	0.0	0.0	not OK

表 7 時系列に沿ってエビデンスを与えたときの X_{10} の条件付確率の変化

与えたエビデンスとその時点	X_{10} の条件付確率	
	OK	not OK
データなし (事前確率)	0.781	0.219
$X_1 \sim X_4$ (入学前)	0.728	0.272
$X_1 \sim X_5$ (入学直後)	0.659	0.341
$X_1 \sim X_6$ (春学期 1/3 経過)	0.566	0.434
$X_1 \sim X_9$ (春学期末)	0.100	0.900

4. 学修支援への活用と今後の展開

本手法を用いることで、学士課程の任意の時点でそれ以降の学修状態を確率として示すことができる。または、未来の状態を仮に入力することで学士課程全体における学修状態の推移をシミュレーションすることも可能である。こうしたモデルの出力によって、介入すべき学生を早期発見したり、学生の気づきを直接促したりすることも考えられる。

実用性を高めるためには優れた汎化能力が必要となるため、ネットワーク構造の決定についてはさらに検討が必要である。

参考文献

- (1) Siemens, G. and Baker, R.S.J.d.: “Learning Analytics and Educational Data Mining: Towards Communication and Collaboration”, Proc. of the 2nd Int’l Conf. on Learning Analytics and Knowledge, pp.252-254 (2012)
- (2) Jayaprakash, S.M., Moody, E.W., Lauria, E.J.M., et al.: “Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative”, Journal of Learning Analytics, Vol.1, No.1, pp.6-47 (2014)
- (3) 植野真臣, “ベイジアンネットワーク”, コロナ社 (2013)
- (4) Weka 3: Data Mining Software in Java: <http://www.cs.waikato.ac.nz/ml/weka/> (参照 2015.6.10)