

# 成績不良の早期予測における機械学習手法の有効性の検証

○安部 恵介 (九州産業大学)

## Effectiveness of Machine Learning Methods in Early Prediction of Poor performance.

\* K. Abe (Kyushu Sangyo University)

**Abstract**— In this paper education support methods utilizing information technology in the university are examined. Machine learning methods are applied to reduce dropouts and postponed graduations by analyzing the data of student's performance and the effectiveness of them in early prediction of poor performance is shown by some examples of practical applications.

**Key Words:** Data analysis, Machine learning

### 1 まえがき

情報化の進展により得られる大量のデータの有効利用が重要な問題となっており、大学においても学内データの収集・分析により、経営改善や学生支援、教育の質向上等に役立てようという意識が高まっている<sup>1)</sup>。

そこで本研究では大学における学内データ活用の概要についてまとめ<sup>2)5)</sup>、特に多くの大学で重要となっている、留年・退学者の防止について検討する。大学においても情報化により、学生の成績や出席状況等の教務データの一元的な収集・管理が可能となってきた。そこでこれらのデータを分析することにより留年や退学を予防する修学支援方式について検討した。

本研究ではまず対象となる学生の教務データを収集・統合し、さらに学生の成績や出席状況等のデータを分析し、機械学習の手法を適用することにより留年や退学を予防する方式について検討する。まず留年・退学しそうな学生を予測し、早期に予防する方式について検討し、さらに日々の出席状況から成績不良となりそうな学生をリアルタイムに予測し、早期に予防する方式についても検討する。

これまでも機械学習の適用について試みたが<sup>6)7)</sup>、本研究ではさらに詳細に分析し、どのような特徴量や手法を用いれば最も効果的かを明確にすることを目的とする。そのためにデータベースとして利用可能なデータ量も増大し、さらに種々の特徴量や手法を適用することにより有効性を検討し、機械学習手法の適用効果を明確にする。

学生には、能力・個性、環境変化など個別の事情があり、定型的データだけでは予測困難な面もある。これらの機械学習によりどの程度予測可能かを明らかにし、個々の学生について詳細に分析することにより、統計的知見と個別の事情を総合的に考慮した支援方式についても検討する。

### 2 大学における学内データの活用

#### 2.1 大学における IR への取組み

大学においても情報化の進展により種々のデータの収集・活用が可能となってきた。これに伴い大学における IR (Institutional Research) への取組みが重要となっている。大学における IR とは大学の経営改善、学生支援、教育の質向上のため、学内データを収集・分析し、改善施策を立案、施策の実行・検証を行うとい

った広範な活動を指す<sup>1)2)</sup>。国内の大学では、学内データ活用による大学改善への期待が高まっており、その具体的実践に向けた取組みが模索されている<sup>1)</sup>。

大学における IR の現状と在り方に関する調査研究も行われており<sup>3)</sup>、全国大学アンケート調査から IR 活動として以下のような項目が挙げられている。

IR 活動の主要な項目：

- ・入学志願者の調査 (マーケティング)、就職状況調査、成績分布や GPA などの分析、休学、留年、中退などの要因分析、学生の達成度調査
- ・学生による大学教育の評価調査など、FD の効果の分析、学生調査 (生活調査、生活実態調査など) 中期計画 (戦略計画) 策定、学生による授業評価の分析、卒業生に対する調査、等。

このような IR に対するニーズ調査を行った結果、データを活用して取り組みたい内容として以下のような項目が挙げられる<sup>1)</sup>。

- (1) 教育の質保証
- (2) 学生支援・キャリア形成
- (3) 入試戦略

#### 2.2 留年・退学者の防止

IR に対するニーズ調査の中で、特に (2) 学生支援・キャリア形成に含まれる「退学者の防止」については大学側が直近の課題としてすぐにでも取り組みたいという強いニーズがある。そのため留年も含めて、留年・退学者の防止に対するデータ分析が注目されてきており、様々な試みが検討されている<sup>1)4)5)</sup>。

<退学者の防止>

- ・AO 入試、推薦入試、一般入試等、入学形態別の分析
  - ・学部・学期ごとの退学傾向
  - ・講義ごとの成績と退学傾向
  - ・GPA と欠席率による退学者予測
- <留年してしまう学生の早期発見>
- ・取得単位数、GPA 等による検出

### 3 教務データ分析による修学支援方式

#### 3.1 学生の教務データ

本学でも大学運営における情報化が進み、多くの学内データの情報システム化による一元的な管理が行われるようになってきた。学生の成績・出席等の教務データもこれらに含まれており、最近では IC カードによる出席管理システムも導入されている。

そこで本研究ではこれらの教務データを活用して学

生の留年・退学の抑制に役立てることを目的とする。学生の成績、履修状況、出席状況等の教務データとして各学生ごとに以下のような情報が得られる。

(1)基本情報

- ・在籍状況:入学、卒業、休学、退学等の年月日
- ・年度別修得単位:各年度前後期の登録・修得単位数、GPA(Grade Point Average)。

(2)履修情報

各年度前後期の履修登録した科目。なお科目には担当教員、単位数、必修・選択区分等の情報も含まれる。

(3)成績情報

各年度前後期の履修登録した科目の評価。評価はS(秀)、A(優)、B(良)、C(可)、D(不可)の5段階で不可の場合は単位を修得できない。またこれらの評価に点数(S:4, D:0等)を与えGP(Grade Point)とし、その平均をGPAとよぶ。GPAは総合成績に対応する。

(4)出席情報

- ・出欠情報:各年度前後期の履修登録した科目の出欠回数。学期ごとの集計値。
- ・出欠情報詳細:各年度前後期の履修登録した科目の毎回の講義ごとの出欠状況。ICカードによる出席管理システムにより毎日更新される。

3.2 修学支援方式

近年、特に退学・留年の増加が問題となっている。そこで教務データを分析することにより退学や留年を予防する修学支援方式について検討した。

図1に教務データ分析による修学支援方式の構成を示す。まず留年・退学防止のために早期予測による指導や履修計画の支援について検討した。例えばA0入試や推薦入試の場合、一般入試に比べて学力が低い傾向にあるが、このような入学時の情報も用いることによりさらに早い時点での学生の傾向を予測することが可能となる。また日々のリアルタイムな出席状況から出席不良への変化をいち早く検出し、成績不良を未然に防ぐことも重要である。そのため全科目の平均出席率を用いて毎週予測を行い成績不良を早期発見する方式についても検討した。

また履修科目の可否予測により成績不良学生の履修計画を支援する方式も重要である。さらに科目分析により履修計画を支援し、修得単位数やGPA等の成績向上を図るとともに、就職との関連も分析することによりキャリア支援に活用することも考えられる。

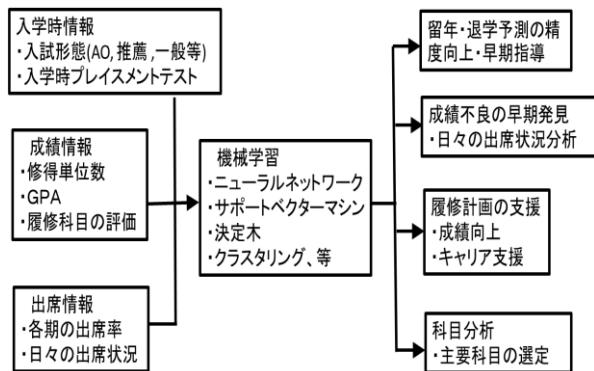


図1. 教務データ分析による修学支援方式

4 成績不良の早期予測における機械学習の適用

学生の成績や出席状況等のデータを分析し、機械学習の手法を適用することにより留年や退学を予防する方式について検討する。

これまでも機械学習の適用については試みたが、本研究ではさらに詳細に分析し、どのような特徴量や手法を用いれば最も効果的かを明確にすることを目的とする。そのためにデータベースとして利用可能なデータ量も増大し、さらに種々の特徴量や手法を適用することにより有効性を検討する。

4.1 留年・退学者の予測

留年・退学しそうな学生を予測し、早期に予防することが重要である。そこで各時点で機械学習の手法を適用することにより、どの程度予測できるか検討した。

入学時では入試区分(AO、推薦、一般)、数学プレジメンテスト成績を用いて予測した。結果を表1に示す。なお一般に機械学習における予測性能は、再現率と適合率(精度)のバランスが重要であり、両者の調和平均であるF値で評価するのが良い。ナイーブベイズの結果が若干良好だが、いずれにしてもこの時点で予測するのは困難である。

次に1年前期終了時では特徴量としてさらに単位取得率と平均出席率も追加して予測した。表2に結果を示す。ここでも留年・退学者のF値で評価すると、ナイーブベイズが若干良好である。しかし、入学時と比べると全般に大きく改善されており、時期や特徴量による差が大きい。

表1. 各手法の比較(入学時)

手法	正解率	分類	再現率	適合率	F値
ニューラル	0.604	留年・退学	0.261	0.481	0.338
		正規卒業	0.821	0.637	0.717
ネットワーク	0.593	留年・退学	0.197	0.444	0.273
		正規卒業	0.844	0.624	0.717
ロジスティック	0.601	留年・退学	0.134	0.452	0.207
		正規卒業	0.897	0.62	0.734
回帰	0.604	留年・退学	0.07	0.435	0.121
		正規卒業	0.942	0.615	0.744
サポートベク	0.65	留年・退学	0.373	0.576	0.453
		正規卒業	0.826	0.675	0.743
ターマシン	0.642	留年・退学	0.373	0.558	0.447
		正規卒業	0.813	0.672	0.735
ランダム					
フォレスト					
ナイーブ					
ベイズ					
最近傍法					

表2. 各手法の比較(1年前期終了時)

手法	正解率	分類	再現率	適合率	F値
ニューラル	0.778	留年・退学	0.585	0.79	0.672
		正規卒業	0.901	0.773	0.832
ネットワーク	0.775	留年・退学	0.577	0.788	0.667
		正規卒業	0.901	0.77	0.831
ロジスティック	0.778	留年・退学	0.606	0.775	0.68
		正規卒業	0.888	0.78	0.83
回帰	0.759	留年・退学	0.613	0.725	0.664
		正規卒業	0.852	0.776	0.812
サポートベク	0.808	留年・退学	0.676	0.8	0.733
		正規卒業	0.892	0.812	0.85
ターマシン	0.764	留年・退学	0.514	0.811	0.63
		正規卒業	0.924	0.749	0.827
ランダム					
フォレスト					
ナイーブ					
ベイズ					
最近傍法					

さらに1年終了時では前期・後期の単位取得率、平均出席率、GPAを用いて予測した。結果を表3に示す。全般にナイーブベイズが良好な結果が得られているが、手法による差はそれほど大きくはない。1年終了時で、留年・退学者を70%以上予測できる。

入学時から2年終了時までの各時点における予測結果を表4に示す。時間が経過するにつれ、予測精度は向上するが、留年・退学を削減するには早期に予防することが重要である。ちなみに2年終了までには8割以上予測可能である。また本学部の規定では3年終了時までには100単位以上取得していないと卒業研究に着手できないため卒業延期となる。実際これにより留年する学生も多い。そこで3年終了時までの取得単位数も予測した結果を表5に示す。

このように複数の手法や指標を総合的に判断することによりほぼ予測は可能であり、留年・退学者の早期予測・予防には十分効果的であることが分かった。なおこれらの予測では検出できない学生が1割程度いるが、これらの学生について教務システムの蓄積情報(個別面談等も含む)を詳細に検討したところ、ほとんどが、能力・個性面、環境変化など個別の事情があることが分かった。

#### 4.2 出席状況に基づく成績予測

日々の出席状況から成績不良となりそうな学生を予測し、早期に予防する方式についても検討した。各学期で取得単位数が10単位以下の学生は留年・退学する割合が多く要注意である。そこで取得単位数が10単位以下を成績不良(I)、10~19単位を中間(II)、20単位以上を成績良好(III)と3つの区分に分類する。

ICカードで出席管理されているので、日々の出席状況をリアルタイムで知ることができる。そこで毎週の全科目の平均出席率を用いて成績予測を行い、成績不良となる学生を予測する方式についても検討した。

通常各科目の授業は1学期に15回(15週)行われる。あまり遅くなってから成績不良予測を行っても、実際には手遅れとなる場合も多い。そこで一例として、まだ修復可能な第6週の時点で、どの程度成績不良学生を予測できるか検討した。

1年前期第6週終了時点での予測結果を表6に示す。特徴量としては、毎週の全科目平均出席率(1週間単位)と、全体の累計平均出席率(6週間分)を用いた。表6に結果を示す。各区分(Iが成績不良)の予測結果の混合行列であり、列が実績、行が予測結果に対応している。

F値で評価する。この例ではランダムフォレストが比較的良好だが、いずれにしてもあまり精度は良くない。1年前期は学生も入学したばかりで、比較的真面目に出席している。後半に入り、最後の方で出席不良となり単位取得できなくなる学生が多いが、この時点ではまだ予測できないためである。

同様に1年後期第6週終了時点での予測結果を表7に示す。こちらはかなり改善されている。1年後期になると学生の出席傾向も大体固まってくるため、成績不良の学生も早期予測が可能となる。さらに前期の単位取得率、平均出席率といった情報も用いることにより(表8)、若干改善されるが、やはり現時点での出席状況が最も重要と分かる。手法としてはニューラルネットワークが若干良好だが、大差はない。1年後期

表3. 各手法の比較(1年終了時)

手法	正解率	分類	再現率	適合率	F値
ニューラル	0.797	留年・退学	0.662	0.783	0.718
ネットワーク		正規卒業	0.883	0.804	0.842
ロジスティック	0.816	留年・退学	0.669	0.826	0.739
回帰		正規卒業	0.91	0.812	0.858
サポートベク	0.811	留年・退学	0.669	0.812	0.733
ターマシ		正規卒業	0.901	0.81	0.854
ランダム	0.811	留年・退学	0.648	0.829	0.727
フォレスト		正規卒業	0.915	0.803	0.855
ナイーブ	0.814	留年・退学	0.725	0.78	0.752
ベイズ		正規卒業	0.87	0.833	0.851
最近傍法	0.797	留年・退学	0.606	0.827	0.699
		正規卒業	0.919	0.785	0.847

表4. 各時期における予測結果

時期	正解率	分類	再現率	適合率	F値
入学時	0.65	留年・退学	0.373	0.576	0.453
		正規卒業	0.826	0.675	0.743
1年前期 終了時	0.808	留年・退学	0.676	0.8	0.733
		正規卒業	0.892	0.812	0.85
1年終了時	0.814	留年・退学	0.725	0.78	0.752
		正規卒業	0.87	0.833	0.851
2年前期 終了時	0.851	留年・退学	0.707	0.87	0.78
		正規卒業	0.937	0.843	0.887
2年終了時	0.887	留年・退学	0.752	0.924	0.829
		正規卒業	0.964	0.870	0.915

表5. 3年終了時での総単位数予測

手法	正解率	分類	再現率	適合率	F値
ニューラル	0.792	100単位未満	0.813	0.654	0.725
ネットワーク		100単位以上	0.781	0.892	0.833
線形回帰	0.797	100単位未満	0.829	0.658	0.734
		100単位以上	0.781	0.9	0.836
最近傍法	0.814	100単位未満	0.789	0.698	0.74
		100単位以上	0.826	0.885	0.855
ランダム	0.816	100単位未満	0.813	0.694	0.749
フォレスト		100単位以上	0.818	0.896	0.855

表6. 1年前期第6週

手法	区分	I	II	III	適合率	F値
ニューラル	I	5	5	0	0.5	0.25
	II	24	32	43	0.323	0.4
	III	1	24	240	0.906	0.876
	再現率	0.167	0.525	0.848	0.741	0
線形回帰	I	13	4	0	0.765	0.553
	II	16	29	35	0.362	0.411
	III	1	28	248	0.895	0.886
	再現率	0.433	0.475	0.876	0.775	0
最近傍法	I	6	3	0	0.667	0.308
	II	21	22	32	0.293	0.324
	III	3	36	251	0.866	0.876
	再現率	0.2	0.361	0.887		
ランダム	I	15	7	0	0.682	0.577
	II	13	23	44	0.288	0.326
フォレスト	III	2	31	239	0.879	0.861
	再現率	0.5	0.377	0.845		

第6週終了時点での単位取得率の実績と予測結果の散布図を図3に示す。

また参考までに、最終的に授業が終了した第15週終了時点での予測結果を表9に示す。ニューラルネットワークによる予測結果であり、1年前期。後期ともに85%程度の精度で成績不良者を予測できる。なお残りの誤差は、やはり出席だけでは予測できない面と、最後の方で急に来なくなり、試験も受けないような予測困難な状況もあるためと考えられる。

1年前・後期における授業開始後3週間ごとに予測した結果を図3に示す。成績不良者の早期発見が目的なので、区分Ⅰ（成績不良）の予測結果のF値の推移を示している。1年前期は入学しばかりで最初の頃は真面目に出席しているが、後半から来なくなる学生がいるため、前半の予測精度はあまり良くない。一方後期になると傾向も安定してき、前期の成績・出席データも利用するため比較的早い時点から予測精度も良好である。

表7. 1年後期第6週

手法	区分	I	II	III	適合率	F値
ニューラルネットワーク	I	47	10	1	0.81	0.764
	II	15	75	47	0.547	0.568
	III	3	42	134	0.749	0.742
	再現率	0.723	0.591	0.736		
線形回帰	I	40	5	1	0.87	0.721
	II	24	81	53	0.513	0.568
	III	1	41	128	0.753	0.727
	再現率	0.615	0.638	0.703		
最近傍法	I	43	5	1	0.878	0.754
	II	20	71	41	0.538	0.548
	III	2	51	140	0.725	0.747
	再現率	0.662	0.559	0.769		
ランダムフォレスト	I	45	10	2	0.789	0.738
	II	19	70	52	0.496	0.522
	III	1	47	128	0.727	0.715
	再現率	0.692	0.551	0.703		

表8. 1年後期第6週（前期情報利用）

手法	区分	I	II	III	適合率	F値
ニューラルネットワーク	I	49	10	1	0.817	0.784
	II	14	74	39	0.583	0.583
	III	2	43	142	0.759	0.77
	再現率	0.754	0.583	0.78		
線形回帰	I	39	4	0	0.907	0.722
	II	24	82	50	0.526	0.58
	III	2	41	132	0.754	0.739
	再現率	0.6	0.646	0.725		
最近傍法	I	43	6	0	0.878	0.754
	II	20	67	37	0.54	0.534
	III	2	54	145	0.721	0.757
	再現率	0.662	0.528	0.797		
ランダムフォレスト	I	45	10	1	0.804	0.744
	II	19	72	51	0.507	0.535
	III	1	45	130	0.739	0.726
	再現率	0.692	0.567	0.714		

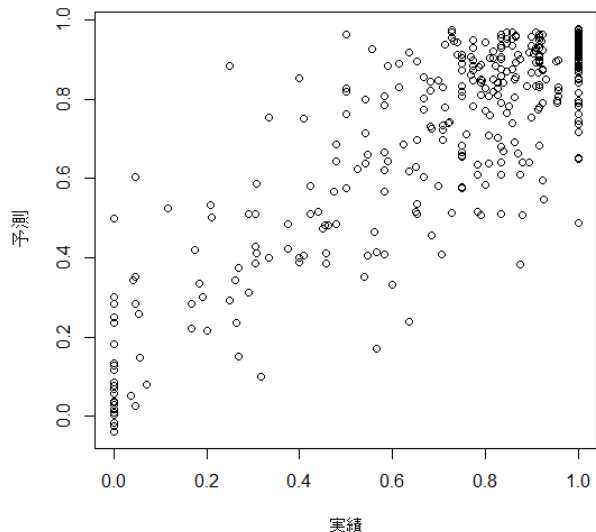


図2. 1年後期第6週での予測結果

表9. 第15週終了時での予測結果

時期	区分	I	II	III	適合率	F値
1年前期	I	24	3	0	0.889	0.842
	II	6	31	42	0.392	0.443
	III	0	27	241	0.899	0.875
	再現率	0.8	0.508	0.852		
1年後期	I	54	7	0	0.885	0.857
	II	10	82	52	0.569	0.605
	III	1	38	130	0.769	0.741
	再現率	0.83	0.646	0.714		

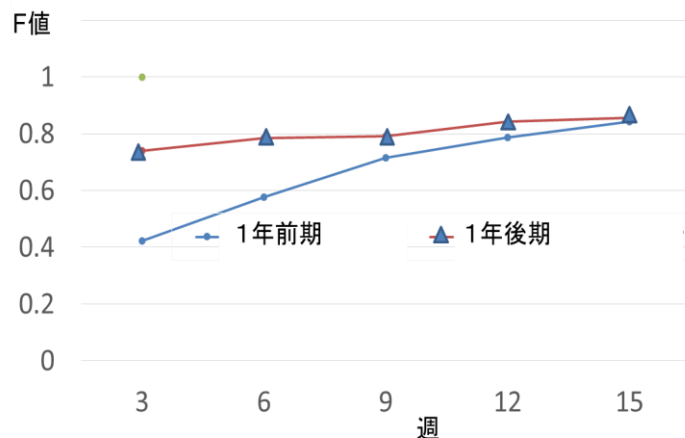


図3. 成績不良者の予測結果の推移



### 4.3 成績不良変化の早期検出

出席状況から成績を予測することにより、成績不良者の早期発見が可能であることが確かめられた。一方入学当初はほとんどの学生が真面目に出席しているが、途中から何らかの理由により出席不良となり、その結果成績不良、留年・退学へと繋がっていく場合が多い。

したがって、このように出席不良・成績不良の状況に陥る学生をできるだけ早期に検出することが特に重要である。そこで出欠状況を図4のようにパターン化し、出欠パターンから成績不良への変化を検出する方式についても検討した。

必修科目の3日間の出欠パターンに畳み込みニューラルネットを適用し成績不良への変化パターンを検出することにより、これまでよりさらに早く成績不良者を検出できることが確認された。また単に出席を白(1)、欠席を黒(0)といった白黒パターンだけでなく、全科目の平均出席率も用いて0～1のグレーの濃淡パターンで表現することにより、さらに検出精度が向上した。

前節で説明した、全科目の平均出席率等の出席状況による成績予測と、今回の主要科目の出欠パターン認識による成績不良変化パターン検出の両者を併用することにより、正確な成績予測と成績不良変化の早期検出が可能となり実用性が向上する。

### 4.4 考察

機械学習により、留年・退学の予測や各学期の途中の時点でどの程度成績不良を予測できるか確かめられた。これにより留年・退学者や成績不良者の早期予防に役立てることができる。各手法の性能評価も行い、有効性も確かめられた。なお実際にはどれか一つの手法に限定するよりは、複数の手法や指標の結果を総合的に判断する方がボーダーラインの学生の検出も可能となり、より適切な支援が行える。

データ量を増やして学習すればある程度予測精度は向上するが、限界はあり、予測と実績の誤差が大きい学生もいる。特に予測結果が大きくはずれるような学生について、個別の面談結果などの情報も用いて詳細に分析すると、能力、個性、環境変化など個別の事情や予測困難な出来事がある場合が多いことが分かった。したがって予測精度の向上だけでは限界があり、機械学習による予測結果という統計的知見とこれらの個別の事情を総合的に考慮して支援する方式が重要である。

予測結果を標準的傾向と考え、誤差の大きい学生は例外的なケースとしてさらに詳細に要因を分析することにより、個々の学生の状況に応じた対応に活用できる。またこれらの分析により得られた要因も新たに特徴量として追加することにより、さらに機械学習の能力を改善することも可能である。特に成績不良学生の場合は面談が行われるのでその所見情報も活用できる。

## 5 あとがき

大学における情報化による学内データを活用した教育・修学支援方式として、近年特に重要となっている留年・退学の削減を支援する方式について検討した。留年・退学の削減支援方式として、学生の成績・出席等の教務データを用いた機械学習の適用により、留年・退学者の予測方式、学期途中における成績不良者の早期予測方式について検討した。種々の特徴量や手

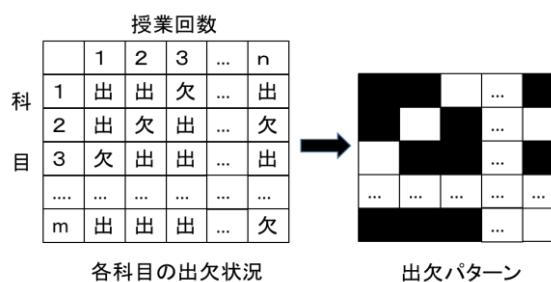


図4. 出欠パターンによる成績不良の検出

法を適用することによりどのような特徴量や手法を用いれば最も効果的かをある程度明確にすることができた。さらにこれらの複数の手法や指標を総合的に判断することにより実用的な方式を構築することができる。

なお今回用いた成績・出席等の情報は時間的に推移するものであり、これを時系列として扱うことによりさらに方式を改善することも考えられる。また今回出欠パターン認識に畳み込みニューラルネットワークを適用したが、このような深層学習の適用によりさらに高精度な手法を検討することも重要である。

また機械学習による予測結果を標準的傾向として定量的に活用することにより、個々の学生の状況により適切に対応することも可能となると思われる。またこれまでの学生との面談記録も所見情報として残されており、これらの情報から新たな特徴量を追加し、改善することも有効である。

さらに多くのデータの中から深層学習により自動的に有用な特徴量を抽出することにより、予測性能を向上させる方式等についても検討していきたい。

## 参考文献

- 1) 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).
- 2) 沖清豪・岡田聡志：データによる大学教育の自己改善—インスティテューショナル・リサーチの過去・現在・展望—, 学文社, 2011.
- 3) 文部科学省高等教育局大学振興課大学改革推進室：大学における IR (インスティテューショナル・リサーチ)の現状と在り方に関する調査研究. [http://www.next.go.jp/a\\_menu/koutou/itaku/1347631.htm](http://www.next.go.jp/a_menu/koutou/itaku/1347631.htm)
- 4) 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).
- 5) 篤田 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).
- 6) 安部恵介：大学における情報化を活用した教育・修学支援方式, 電気学会電子・情報・システム部門大会講演論文集, pp. 691-696, 2016年8月.
- 7) 安部恵介：大学における学生データの機械学習を用いた分析による修学・教育支援方式の検討, 計測自動制御学会知能システムシンポジウム講演論文集, A1-1, pp. 1-6 (2018).