データアナリティックスを用いる大学教育支援環境の検討

豊 川 和 治

Kazuharu Toyokawa. A Study of Academic Adviser Support by Analytics of Student Data. *Studies in International Relations* Vol.36, No.1. October 2015. pp.79 − 86.

An academic adviser is required to guide a student to take academic courses and also to give advises to improve and encourage his/her learnings. This work is to propose a new information system capable to support academic advisers. Using multiple discriminant analysis of students' academic activities, the system can provide information on student advantages and risks related to course learnings. Applying the system to actual activities of an academic adviser, provided information is proved to be useful to help students improve his/her learning.

1. はじめに

近年，大学への進学率の上昇，入学試験制度の多様化，海外からの留学生の増加などにより，学習履歴や習熟度の異なった学生が，共に大学で学ぶようになった。このような様々なバックグラウンドを持つ学生が，それぞれカリキュラムを中から適切に科目を選択して履修し，幅広い教養と，専門的な技能や知識を身に付けることができるよう，大学が支援することが求められている[(1)]。

我々は，以前より大学に在籍する学生各々に，教員のアカデミック・アドバイザーを付け，適時，科目履修，GPA（Grade Point Average）に基づく履修に関する指導を行い，特に最終学年の学生には，必要に応じて卒業指導を行ってきた。

しかし，学生の中には4年間の修学で卒業要件を満たすことができず，退学または卒業延期という学生が一部存在する。この状況を改善するためには，就学年数の低い段階から，学生への積極的な履修指導を行うことが必要となる。

アドバイザーが学生と相談するために必要な情報を提供する従来のシステムは，履修登録，履修成績，単位取得状況などの，個々の学生の学習情報を参照するシステムだった。修業年数の低い段階では，履修単位数など，限られた情報だけで，履修指導を行わなければならない弱点があった。

我々は先に，学務情報を中心に学生に関して入手できる情報を集積し，データアナリティックスを用いて情報を分析し，得られた学生の学修到達度に関する知識を用いて，学生に対する卒業指導，就職活動などのアドバイスを支援するシステムを提案した[(2)]。

今回はこの手法をさらに発展させ，学生の就業年数の各段階で，修学に関するアドバイスを行うための情報提供を支援するシステムを検討する。

この報告では，2節で学生に関する統合データベースの構築の概要と，データ分析方法について述べ，3節ではデータの主な分析結果，得られた知見について述べる。4節では，このシステムを学生へのアドバイスに応用し，システムの有効性を検証すると共に，今後の研究の見通しについて述べる。

2. 学生に関するデータの収集

分析に用いた学生のデータは，学籍番号，学科，性別，出身地，出身校，入試種別などの学籍情報，修得単位数の累計，成績評価でS，A，B，Cを得た科目数，GPAなどの成績情報，履修した科目，配当単位数，科目成績，修得単位，履修年度，学期などの履修情報，休学・退学，卒業情報などである。これらの学生データと，属性項目をFig.1に示す。

学籍情報

キー	属性項目	インスタンス
学籍番号	学年	1，2，3，4
	学科	A，D，E，F
	性別	男，女
	出身地	静岡県，神奈川県，東京都，・・・その他
	出身校	高校名
	入試種別	一般入試1期，2期，推薦，AO・・
	住所区分	自宅，自宅外

成績情報

キー	属性項目	インスタンス
学籍番号	累計修得単位数	単位数
	科目成績 S	取得科目数
	科目成績 A	↑
	科目成績 B	↑
	科目成績 C	↑
	取得科目合計	↑
	GPA	0.00～4.00

履修情報

キー	属性項目	インスタンス
学籍番号	科目名	国際関係論Ⅰ，・・・・
	単位	2
	科目成績	S，A，B，C，D，E，N，P
	習得単位	
	年度	2010，・・・，2014
	学期	前期／後期
	教員氏名	・・・

卒業情報

キー	属性項目	インスタンス
学籍番号	卒業年	年・月

退学情報

キー	属性項目	インスタンス
学籍番号	退学・除籍	退学／除籍
	退学・除籍年	年・月・日

Fig.1　分析に使用したデータ項目

これらのデータは，学務情報システムで管理されている。今回のデータ分析のため，2009～2014年度の学期末時点の学部在籍学生の内，3学科，2・3・4年次の合計368名分のデータセットを使用した。個人情報の保護，教育指導のための使用という本来の情報利用目的遵守の確認，情報の保管とセキュリティについては細心の注意を払った。

これらのデータより，学生の学期ごとの教科の履修，その積み重ねの結果，卒業に至るか，又は卒業延期（留年），退学に至る過程を観察するため，各学生の各年次・学期末までの履修データを集計したものを分析の対象のデータとした。

各学生の履修に関わるものとして採用した変数は，履修の進行状況を表す量的変数群：修得単位数，成績評価S，A，B，Cを受けた科目数，平均点，各学期及び累積GPAである。4年間の履修の結果，修了，留年，或いは退学は，質的な変数：卒留退として扱う。

学生のプロフィールに関わる質的変数群として採用したものは，性別，入試種別，住所区分データである。これ以外の変数の出身地，出身校は，変化が多岐にわたり，精度を持って統計量を算出できる標本数を確保できないため採用しなかった。

採用した質的変数は，Fig.2のようにダミー変数として数値化した[1]。

変数名	区分	数値コード
卒留退	卒業	0
	留年	1
	退学	2
性別	女	0
	男	1
通学区分	自宅	0
	自宅外	1
入試種別	指定校＆公募制	0
	留学生	1
	一般入試（第2期）	2
	一般入試（第1期）	3
	センター入試	4
	特別推薦	5
	AO	6
	推薦・選抜	7
	保体審	8

Fig.2　質的変数の数値化ルール

4年間で，所定の課程を修了する学生（卒業生）に比較して，卒業延期，退学を余儀なくされる学生の各年度，学期の修得単位数・GPAを観察すると，Fig.3に示すように，比較的順調に履修を続けていたものが，ある学期に突然，習得単位数が極端に少なく（例えば10単位未満），かつGPAが良くない（例えば1未満）学期が認められることが多い。

そこで，この学期を「失策」の学期として，これまでの在籍学期数のうち，失策学期の数をパー

セント表示し失策率として，分析対象の変数に加えた。

卒業，退学情報については，収集した学生情報の内，2014年度末（2015年3月）までに4年在学し，卒業，留年，あるいは退学したかどうかで変数：卒留退に数値コード0～2を与えた。2014年度末でまだ1～3年次の学生に対しては，空欄とした。

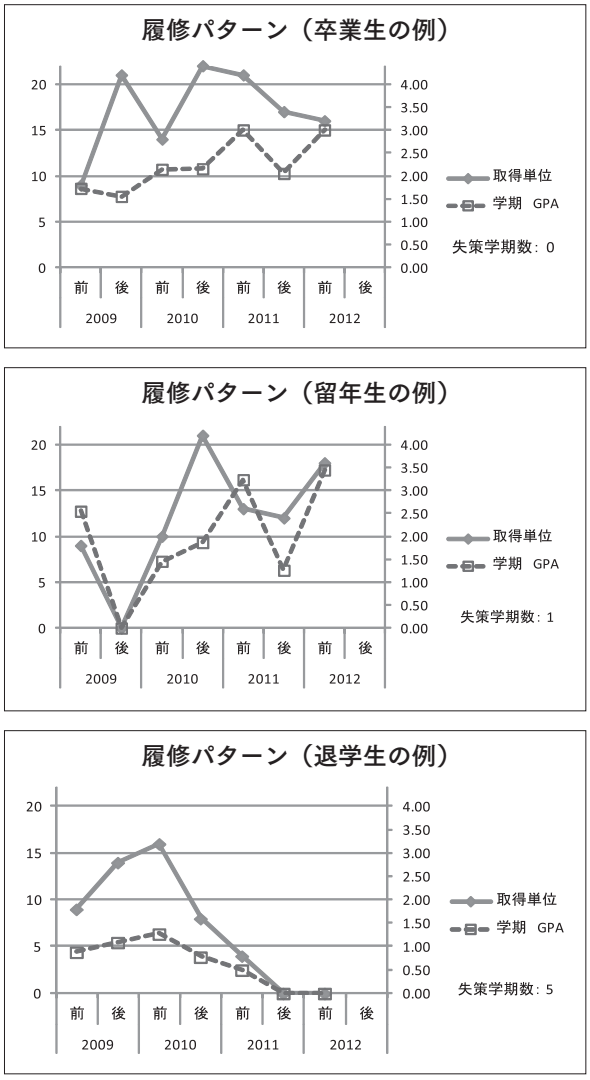

Fig.3　履修パターンの例

このデータから，どのような知見が得られるか調べるため，統計的な分析を行った。分析には，SPSS PASF 23-J（Windows版）及びIBM Modeler 12.0（Client版）を使用した。

3. 退学・卒業延期リスクの分析

3.1. 退学・卒業延期と相関のある変数

学部学生の大半は，4年間の修学で卒業要件を満たす単位を修得し卒業するが，これを満たすことができない一部の学生は留年，また退学の道を選ぶ。分析の目的は，この留年，退学のリスクを，得られた学生情報より，統計的に信頼できる形で予測することである。ここでは，回帰分析の一種の判別分析を用いた[(3)]。

まず，学生の4年間の学修結果を表す変数：卒留退と有意な相関のある変数を調べた。この結果をFig.4に示す。結果変数：卒留退と統計的に有意な相関のある変数として，科目単位数，失策率，累積GPA，成績評価S，A，B，Cの数，性別の8変数を説明変数として選んだ[2]。

		卒0留1退2	
科目単位数	Pearsonの相関係数	－0.870	**
	有意確率（両側）	0.000	
平均点	Pearsonの相関係数	－0.474	**
	有意確率（両側）	0.000	
評価Sの数	Pearsonの相関係数	－0.481	**
	有意確率（両側）	0.000	
評価Aの数	Pearsonの相関係数	－0.630	**
	有意確率（両側）	0.000	
評価Bの数	Pearsonの相関係数	－0.407	**
	有意確率（両側）	0.000	
評価Cの数	Pearsonの相関係数	－0.003	
	有意確率（両側）	0.971	
累積GPA	Pearsonの相関係数	－0.682	**
	有意確率（両側）	0.000	
性別名称 男子学生：1 女子学生：0	Pearsonの相関係数 有意確率（両側）	0.229 0.011	*
通学区分名称 自　宅：0 自宅外：1	Pearsonの相関係数 有意確率（両側）	0.101 0.268	
入試種別名称	Pearsonの相関係数	0.113	
	有意確率（両側）	0.216	
卒留退	Pearsonの相関係数	1	
	有意確率（両側）		
失策率（%）	Pearsonの相関係数	0.818	**
	有意確率（両側）	0.000	

** 相関係数は1%水準で有意（両側）です。
* 相関係数は5%水準で有意（両側）です。
データ　D学科4年次前期末（122名）

Fig.4　退学・卒業延期との相関のある変数

他の変数，例えば通学区分名称，入試種別名称は，得られたデータでは有意な相関が認められなかったので，判別分析の説明変数から除外した。また平均点は累積GPAとデータの変化が似通っていて，冗長性があるので採択しないこととする[3]。

3.2. 退学・卒業延期の予測モデル

4年間の修業で無事卒業できる学生は，留年や退学を余儀なくされる学生とは，各年次の就業パターンに明確な違いあるという仮説を検証することを試みる。

先に収集した学生データは，1年次から入学した学生（本科生）と3年次に短期大学から編入した学生のデータを含む。編入学生に関しては，履修データ中には1，2年次の成績評価情報がなく，認定された単位数だけ存在する。また，3年次に履修できる単位数制限値も本科生より大きい等，制度上履修パターンに違いが生じると考えられるので，本科生，編入生に関しては，以下で別々に履修パターンを判別分析するモデルを設定する。

本科生データのうち，2012年度A学科，D学科の4年生183名のデータから，卒業留年退学を判別する履修パターン学習用データとして141名分，その判別を検証するデータとして42名分をとり分ける。

はじめに3年次学年末時点の学習用データで，4年次学年末の，卒業＝0，卒業延期＝1，退学＝2と結果変数を設定し，判別分析することにより判別モデルを学習する[4]。

学習した判別モデルを用いて，取り分けておいた検証用データを用い，判別精度を検証した。

Fig.5に示すように，学習データに対しては88.8%，検証データに対しては95.2%の判別精度が得られた[5]。

学習データ分類結果a

		卒0留1退2	予測グループ番号			合計
			0	1	2	
元のデータ	度数	0	103	4	0	107
		1	5	22	6	33
		2	0	2	10	12
	%	0	96.3	3.7	.0	100.0
		1	15.2	66.7	18.2	100.0
		2	.0	16.7	83.3	100.0

a　グループ化された個数の内88.8%が正しく識別された

検証データ分類結果b

		卒0留1退2	予測グループ番号			合計
			0	1	2	
元のデータ	度数	0	35	0	0	35
		1	1	3	0	4
		2	0	1	2	3
	%	0	100.0	.0	.0	100.0
		1	25.0	75.0	.0	100.0
		2	.0	33.3	66.7	100.0

b　グループ化された個数の内95.2%が正しく識別された

Fig.5　判別モデルの学習と検証結果
（A，D学科本科4年生　3年次末予測）

ではこの判別モデルは，より就学年数の低い段階でどの程度正しい予測ができるか検証を行った。結果をFig.6に示す。

	学習データ	検証データ
1年 学年末	83.7%	76.2%
2年 学年末	78.7%	85.7%
3年 学年末	88.8%	95.2%
4年 前期末	90.1%	95.0%

Fig.6　本科学生の修業年次ごとの判別精度

識別に用いたA，D学科4年生141名のデータでは，4年次学年末で，卒業104名（73.8%），卒業延期31名（22.0%），退学6名（4.3%）だった。例年ほとんどの学生は4年で卒業する筈なので，今年全員卒業という楽観的な仮定をすれば，判別精度は約74%ということになる。

1年次でもこの楽観的仮定を若干上回る76%～84%の判別精度が得られていることは，初年度の学業データから3年後をこの判別モデルはある程度予測する知見を得ていることがわかる。

編入学生に関しては，2009年から2011年にA，B，C，D学科に3年次編入した学生データ47名

分から，履修パターン学習データ30名分，検証データ17名分をランダムに選択して取り分けた。このデータの中には，4年次末迄に退学した学生はなかった。

判別分析の結果をFig.7に示す。

学習データ分類結果a

		卒0留1退2	予測グループ番号		合計
			0	1	
元のデータ	度数	0	25	0	25
		1	1	4	5
	%	0	100.0	.0	100.0
		1	20.0	80.0	100.0

a　グループ化された個数の内96.7%が正しく識別された

検証データ分類結果b

		卒0留1退2	予測グループ番号		合計
			0	1	
元のデータ	度数	0	12	1	13
		1	0	4	4
	%	0	92.3	7.7	100.0
		1	.0	100.0	100.0

b　グループ化された個数の内94.1%が正しく識別された

Fig.7　判別モデルの学習と検証結果
（A，B，C，D学科編入生3年次末予測）

編入生は在学期間が短いので，各学期毎に卒業・留年の判別予測を行った。結果をFig.8に示す。

	学習データ	検証データ
3年 前期末	93.3%	94.1%
3年 学年末	96.7%	94.1%
4年 前期末	100.0%	88.2%

Fig.8　編入学生の修業学期ごとの判別精度

3.3. 学習データ選定による予測のぶれ

学習した判別モデルは，学習するデータにより，どのように判別精度が変化するか調べる。

学習データとして，2012年度A学科4年生61名，2012年度D学科3年生164名をそれぞれ選定し，学習した判別モデルを，先と同じ検証データ2012年度D学科42名分に適用した。得られた判別精度をFig.9，Fig.10に示す。

学習データに対する判別精度は90.9%から91.8%と高いが，検証データに対する判別精度は，83.3%～92.9%と開きを生じている。

学習データ分類結果a

		卒0留1退2(+)	予測グループ番号			合計
			0	1	2	
元のデータ	度数	0	49	1	0	50
		1	2	5	2	9
		2	0	0	2	2
	%	0	98.0	2.0	.0	100.0
		1	22.2	55.6	22.2	100.0
		2	.0	.0	100.0	100.0

a　A学科4年生（61名）2年次学年末データによる学習グループ化された個数の内91.8%が正しく識別された

検証データ分類結果b

		卒0留1退2(+)	予測グループ番号			合計
			0	1	2	
元のデータ	度数	0	34	1	0	35
		1	1	3	0	4
		2	0	1	2	3
	%	0	97.1	2.9	.0	100.0
		1	25.0	75.0	.0	100.0
		2	.0	33.3	66.7	100.0

b　グループ化された個数の内92.9%が正しく識別された

Fig.9　判別モデルの学習と検証結果
（D学科4年生　2年次末予測）

学習データ分類結果a

		卒0留1退2(+)	予測グループ番号			合計
			0	1	2	
元のデータ	度数	0	131	3	3	137
		1	3	12	3	18
		2	2	1	6	9
	%	0	95.6	2.2	2.2	100.0
		1	16.7	66.6	16.7	100.0
		2	22.2	11.1	66.7	100.0

a　D学科3年生（164名）2年次学期末データによる学習グループ化された個数の内90.9%が正しく識別された

検証データ分類結果b

		卒0留1退2(+)	予測グループ番号			合計
			0	1	2	
元のデータ	度数	0	33	2	0	35
		1	1	1	2	4
		2	0	2	1	3
	%	0	94.3	5.7	.0	100.0
		1	25.0	25.0	50.0	100.0
		2	.0	66.7	33.3	100.0

b　グループ化された個数の内83.3%が正しく識別された

Fig.10　判別モデルの学習と検証結果
（D学科4年生　2年次末予測）

3.4. 判別モデルの応用

判別モデルを学習すると判別得点の計算式：判別関数が学習により与えられるので，個々の学生に対して判別得点 *(x,y)* が算出されると共に，卒業者グループ，留年者グループ，退学者グループの判別得点のグループ平均（ここで重心と呼ぶ）の位置 *(X,Y)* が導き出される[6]。

Fig.11 に，A，D学科4年生141名の3年次の時のデータに対する，卒業・卒業延期・退学者グループの重心の位置 *(X,Y)*，Fig.12 に，個々の学生の判別得点 *(x,y)*，及び卒業・卒業延期・退学者グループの重心のプロットを示す。

これより個々の学生の得点と，各重心との位置関係から，学生の卒業・留年・退学の確率が予想できる[7]。さらに，学習した判別関数により，学習していない未知の学生に対してもデータが得られれば，判別得点，卒業・留年・退学の確率も算出できる。

卒業延期あるいは退学者の重心の近い位置に個々の学生の判別得点の位置が近いと，卒業延期あるいは退学のリスクが高いと予想できることが分かる。

例えば3年次末時点で，学生データの，Aさん（判別得点 $x=0.885$，$y=0.418$）は卒業確率97.9%で，Bさん（判別得点 $x=-1.16$，$y=-1.73$）は卒業延期確率90.0%だった。追跡調査した結果，1年後の2013年3月の学年末に，Aさんは卒業し，Bさんは留年が確定してしまった。

卒留退	判別得点	
	関数 X	関数 Y
0：卒業	.757	.043
1：卒業延期	−1.817	−.340
2：退学	−3.742	1.006

Fig.11　グループ重心の関数

このことから，学習された判別モデルは，学生の履修に関してアドバイザーが相談を受ける際に，有効な知見を与えると期待される。また問題がある学生の改善の度合いを，その後の学期においてフォローすることにも有益であると期待される。

4. アカデミック・アドバイザー支援への応用

これまで述べた判別モデルによる学生の履修過程の分析から知見を得て，アカデミック・アドバイザーをどのように支援できるか検討する。

任意に選んだ10名の学生（前節の分析に使用したデータの中から5名分，それ以外の未使用のデータから5名分）の1，2年次の分析結果をFig.13に示し，アドバイスの概要，その後の学生の履修状況を追跡調査した。得られた知見を以下に示す。

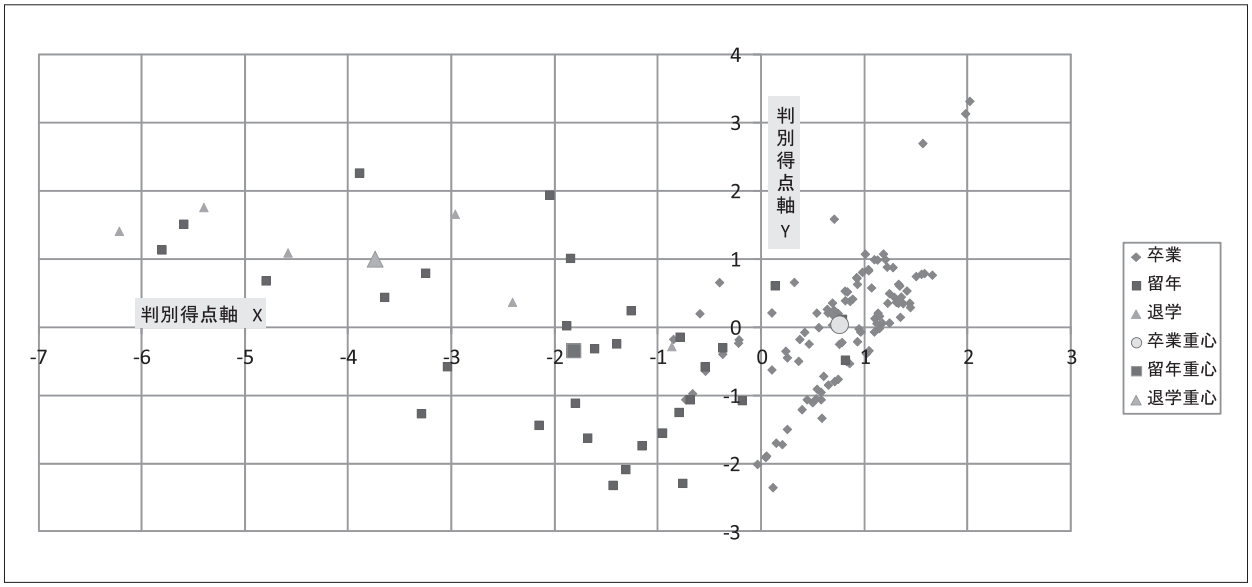

Fig.12　各学生の判別得点プロット（A，D学科本科生141名，3年次末）

学生ID	1年次末				2年次末			
	修得単位数	卒業確率	留年確率	退学確率	修得単位数	卒業確率	留年確率	退学確率
1	12		70.8%	26.2%	44	14.2%	80.8%	
2	18		47.8%	45.4%	37	9.4%	85.4%	
3	31	62.4%	36.8%		63	79.3%	20.6%	
4	32	75.1%	24.7%		72	87.9%	12.0%	
5	32	75.4%	24.3%		70	90.6%	9.4%	
6	36	84.4%	15.4%		68	59.2%	40.5%	
7	37	82.2%	17.5%		76	91.8%	8.2%	
8	39	87.2%	12.6%		79	97.1%	2.9%	
9	40	92.8%	7.1%		80	96.3%	3.6%	
10	40	85.2%	14.6%		69	33.6%	65.6%	

Fig.13　アカデミック・アドバイザー支援情報の抽出例

①　1年次の修得単位数が極端に少ない学生（ID：1，2）は，退学確率が20～50%にもなり，退学の危険がある。

1年次末のとき，学生1はアドバイザーに，専門学校か大学かを迷いつつ大学に入学したこと，現在専門学校に転校も考えていることを話した。アドバイザーは，資格を取得するだけなら専門学校だが，大学は専門教育と，それを支える教養の両方が学べるので，専門性を長く保って社会で活躍するために，大学教育が生きると話した。学生も理解し，在学継続を決意し，2年次から学修に改善がみられた。その結果4年次に無事修了した。

学生2はサークル活動に熱を入れすぎて，授業に身が入らず，アドバイザーが注意したので2年次以降少し改善が見られた。しかし4年次まで学修スピードがさほど上がらず，その結果1年留年して5年目に卒業した。

②　比較的成績の良い学生（ID：3～10）も1年次のとき留年確率は，数%～40%の表示が出る。しかし2年次にこの確率が下がり改善が認められる場合は心配が少ない。

実際，学生3，5，7，8，9は4年次に卒業した。しかし学生4は，予想が外れ半年留年した。

2年次にこの確率が上昇する学生は，4年次に留年する危険性が高い。実際，学生6，10は，1年留年した。

③　従来，学年終了時に一定の取得単位数を満たさない学生に対して，アドバイザーが学生との面談を実施するなど特別ケアを行うシステムがある。しかしこのシステムは，単位数だけで学生の留年・退学リスクを評価するのは無理があると考えられる。

例えば，1年次終了時20単位以下，2年次終了時36単位以下というケア設定ラインを決めると，Fig.14の学生1，2，10に対しては，2年次終了時の設定ラインからはケアの対象外であるが，留年のリスクは非常に高いことが分かり，是非ケアすべき対象であったことがわかる。

このように，判別分析から得られる知見を，学生にアドバイスを実施するのに，学業からのドロップ・アウト防止などの施策に，有益であることが確認された。

今後の発展のため，この学生データの分析の目的を拡大し，在学期間中の海外留学，ダブルデグリー取得，外国語スピーチコンテストへの出場・受賞など優れた学業結果を学生が達成するためのアドバイスなど，より木目の細かい，詳細な分析が必要とされると考えられる。

このためには，正確でより長期にわたる学生データの蓄積と分析が，不可欠であると考えられる。

参考文献

(1)　大学審議会：“大学入試の改善について（答申）”，文部科学省，(2000)；http://www.mext.go.jp/，2015年6月30日アクセス

(2)　豊川 和治：“データマイニングを用いる学習アドバイザー支援システムの検討”，教育システム情報学会研究報告，vol.27，no.1，pp.43-50，(2012-5)．

(3)　竹内 啓：“数理統計学－データ解析の方法”，東洋経済新報社，(1963)．

[1] 質的変数のうち，性別，通学区分は（0，1）の二値で，どちらかを区分する。一方，卒留年，入試種別は，その変数の程度を数値で表す「順序変数」である。入試種別に関しては，与えられたデータセットで，各グループのGPAの平均値で順序を付けた。

[2] 変数：成績評価Cの数は，卒留退と相関が弱いが，評価S，A，Bの数を採択しているので，加えて採択した。回帰分析では，S，A，B，Cの数の合計：総修得科目数を，説明変数に加えているのと同等である。

[3] 回帰分析では，説明変数に特性の似通った変数が混在すると，判別精度の劣化（多重共線性の問題）をもたらす事がある。従って，ここでは，その危険性を避け，平均点と累積GPAは，互いに特性が似通って冗長であるとして，片方の累積GPAのみ採択した。

[4] 判別モデルの学習とは，結果変数（この場合変数：卒留退），及び，説明変数からなる学習データに判別分析を行い，正準判別関数係数を決定することである。得られた係数と各学生の説明変数の線形結合で判別得点が算出でき，卒留退のいずれに最も近いか，確率を含めて判別できる。

[5] 判別精度は以下のように算出できる。例えば，Fig.5の学習データ分類結果では，結果変数：卒留退が，元々‘0’（卒業）であったものを‘0’’と正しく判別したケースが103，元々‘1’（留年）を‘1’と判定したケースが22，元々’2’（退学）を’2’と判定したケースが10であるから，判別精度はこの正解ケースの数を，全ケース152で割り，（103＋22＋10）÷152で88.8％となる。

[6] 判別分析を行うと，一般に，複数個の独立な判別関数が与えられる。今回の判別分析では，2個の独立な判別関数が算出されたので，個々の学生の判別得点を *(x,y)* で，重心を *(X,Y)* と，2次元座標で表現することとする。

[7] 判別分析においては，どのグループに属するかの確率は，各重心を中心とする正規分布に従うと仮定して算出する。この仮定が成り立つことは，Wilksのラムダのχ2乗検定で判定して確認する。

正規分布に従う場合，判別得点がある重心に近いと，その重心に属する確率が1に近くなる。今回，判別得点は2次元であるから，確率はx方向，y方向それぞれ異なった標準偏差に見合う楕円形の広がりをもった正規分布となる。