2019 年 度　　情 報 工 学 科 卒 業 研 究 概 要

学生の生活及び修学データを用いた
ロジスティック回帰分析による要注意学生の推定

ネットワーク系　舟橋研究室
No. 27115132　福田　太一

1　はじめに

教育現場において「将来的に留年および退学する可能性の高い学生」を意味する「要注意学生」[1] の存在が問題視されている. 本研究では要注意学生の早期の抽出を目的に, ある年度の 4 年次の学生 110 人の学生生活実態調査の結果と成績の指標となる GPA を用いたロジスティック回帰分析を行った. なお, これらのデータには個人を特定する情報は含まれていない.

2　データマイニングの手法

2.1　主成分分析

多くの変数の持つ情報を主成分と呼ばれる複数の合成変数に縮約する方法. この主成分に変数の値を入力して求められる数値を主成分得点と呼ぶ.

2.2　変数選択法

ロジスティック回帰モデルの変数選択法に, 全ての説明変数を採用する強制投入法と説明変数を徐々に減らしてモデルを作成するステップワイズ法を採用した.

3　データの詳細

3.1　目的変数

目的変数には学生の留年判定データを採用した.

3.2　説明変数

説明変数には次の 3 種類のデータを採用した.

- GPA： 学生の成績の指標となるデータ.
- 睡眠データ： 学生の平日と休日それぞれの就寝時間, 起床時間, 睡眠時間を記録しているデータ.
- 住居・通学データ： 学生の出身高校所在地, 住居, 住所, 通学時間, 通学手段, 入構手段, 同居人に関する学生生活実態調査のアンケート結果を記録しているデータ.

4　ロジスティック回帰分析による要注意学生の推定

学生のデータを投入したモデルの出力があらかじめ設定した閾値 (確率) を超えた場合にその学生を要注意学生と推定する. 再現率と適合率の調和平均である F 値, 要注意学生かどうかをを正しく予測できた割合である正解率の 2 つを指標にモデルを評価する.

採用する説明変数の組み合わせを 4 通り用意し, それぞれにおける最良モデルをケース 1〜4 とする (表 4.1). これらが算出した正解率, F 値 (表 4.2) を比較した結果, ケース 2 ≥ ケース 1, ケース 3 > ケース 4 となった. ケース 1 がケース 2 より劣る結果から, 要注意学生の推定において住居・通学データがノイズであること, ケース 2 が最も優秀であった結果から要注意学生の推定に最も寄与しているデータの組み合わせは GPA と睡眠データであるということが分かった. また, GPA と睡眠データを採用した場合は入力に実データ, 変数選択にステップワイズ法を用いることでモデルの推定率が最も高くなることも明らかとなった.

表 4.1: ケース (採用データごとの最良モデル) の定義

	採用データ	入力	変数選択
ケース 1	GPA, 睡眠 住居・通学	主成分 得点	ステップ ワイズ法
ケース 2	GPA, 睡眠	主成分 得点	ステップ ワイズ法
ケース 3	GPA 住居・通学	実データ	ステップ ワイズ法
ケース 4	睡眠 住居・通学	実データ	強制 投入法

表 4.2: 各ケースの精度評価

	閾値	正解率	F 値
ケース 1	20%	1.00	1.00
	1.8%	0.83	0.17
ケース 2	20%	1.00	1.00
	1.8%	0.90	0.27
ケース 3	20%	0.98	0.67
	1.8%	0.86	0.20
ケース 4	20%	0.95	0.40
	1.8%	0.41	0.06

5　むすび

学生の生活データと修学データを用いたロジスティック回帰分析の結果, 要注意学生の推定において住居・通学データは結果にノイズをもたらし, GPA と睡眠データは要注意学生の推定に寄与し, モデルの入力と変数選択はそれぞれ主成分得点, ステップワイズ法が有効であるという結論に至った. 更なるモデルの推定率向上に向けた課題として, GPA と睡眠データの関係性の調査と, これら以外のデータを追加して新たなモデルを作成することが挙げられる. 将来的に要注意学生の推定率の高いモデルを作成し, 早期の要注意学生推定・削減に貢献したい.

参考文献

[1] 伊藤圭佑：“ データマイニングによる『要注意学生』の発見に関する研究 ”、平成 25 年度名古屋工業大学修士論文、2013.