ICT 上の学生データを用いた中途退学者の分析手法の検討

Analysis method of dropout students using student ICT-based data

高橋 大樹[*1], 小松川 浩[*1]
Hiroki TAKAHASHI[*1], Hiroshi KOMATSUGAWA[*1]
[*1] 千歳科学技術大学大学院光科学研究科
[*1]Graduate School of Photonics Science Chitose Institute of Science and Technology
Email: takahashi214@kklab.spub.chitose.ac.jp

あらまし：我々は先行研究にて Deep Learning を用いて, 大学内のデータを分析し, 在学生の中から将来的に中途退学する学生を推論するプログラムの開発を行ない, 76%程度の推論精度を得た. 本研究では, これとは別に, データ分析に基づき明示的に退学者の特性を示す特徴量を調べた. この特徴量を入力データの一部として Deep Learning に適用することで, 別の特徴量抽出を行えるか, 適切なネットワーク構造を調べることが可能と考えている.
キーワード：ICT 中途退学者推論 データ分析 プログラム

1. はじめに

学生のデータを活用して退学者動向の解析は,Institutional Research(IR)の観点でも重要なテーマになっている. 2014 年に文部科学省が公表した中途退学者の実態調査の結果では, 1163 校の大学・短期大学・高等専門学校に対し, 2012 年度中途退学者の状況を調査したところ, 同年代における退学者は, 全学生数の 2.65%にあたる 79, 311 人となっている(1). また近年, Deep Learning(以下 DL) が機械学習の新たなアプローチとして注目を浴びている. DL に関する話題は画像認識, 音声認識, 自然言語処理, ゲームなど様々な領域に広がり, 教育への適用も期待されている. そこで本研究では先行研究(2)で行った DL を用いた中途退学者推論の精度とデータを統計を用いて分析した結果を比較し, その推論精度を検証した.

2. 本研究の目的

これまでの分析では, 分析者がそれまでの経験から入力するデータを決定し分析していた. 近年ハード,ソフトウェアの性能向上により, DL を用いて利用可能である全てのデータを入力して, 分析することが可能となった. しかしどのデータが主な特徴として現れているかどうかを把握することが難しかった. そこで DL で特徴量があることが分析できたデータ自体を分析し, そのデータの特徴量を把握することで, DL に入力するデータを設定することが出来, DL の精度の向上を見込むことができると考えた.

3. 先行研究

3.1 概要

先行研究では中途退学者推論プログラムを用いて, 中途退学者の推論精度のさらなる向上を目的として推論を行った. A 大学のデータを学習データとして入力し, 分析を行った.調査の結果, 学習比率を 80:20 とし,中間層が 5 層でノード数が 10 個>15 個>20 個>25 個>30 個と増加していく構造で, batchsize が 548, 学習回数が 1000 回という条件において, 推論精度が 76%程度となった.

3.2 先行研究で利用したデータ

取り扱う学生のデータに関しては, 研究倫理委員会による確認のもと手続きを行い, 学生番号や氏名などの個人情報を特定できる情報を匿名化して取り扱った. 本研究で用いたデータは A 大学の複数のデータから取得し, データセットにした. データセットとは本研究で用いるために整えたデータ群である. 本調査の対象は, 1998 年度から 2017 年度に A 大学へ入学し入学前教育を受けた学生から一部を抽出した計 3514 人となった. 3514 名の内, 中途退学をした学生の数は一定数存在している. 本研究では, E ラーニングシステムと大学の講義を管理するシステムと学生の過去の成績データを管理するシステムの 3 つのデータベースから得た 37 列のデータを使用した. E ラーニングシステムは学習時間,学習の進捗率、E ラーニングを利用したテストの結果などの学習状況のデータを利用し, 大学の講義を管理するシステムでは入学年, GPA, 学科, 出席率などのデータを利用し, 学生の過去の成績データを管理するシステムでは高校のランク, 入学方法, 基礎学力テストの結果などのデータを利用した.

4. 特徴量の抽出

4.1 概要

本研究では先行研究で利用したデータセットを分析して, 退学者の特性を示す特徴量を調べた. まず全ての列からデータの欠損が 7 割を超える列を排除した. その結果 16 列のデータとなった. その後重複して存在する列データを統合し, 8 列のデータとなった. 一連のデータ列の項目を表 1 に示す. 8 列のデータには GPA などの成績情報がなかったため, 新たにデータベースから一年必修の前期科目と後期科目

の成績の二列のデータを取得した．これらの成績情報は数学と情報の成績情報である．その際2008年から2016年のデータだけを抽出し，1864件のデータとなり，退学者数は同じく一定程度の割合で分布している．

表 1 データセット

E ラーニングの学習結果
高校のランク
入学方法
学科
入学年
出席率
数学の成績
情報の成績

4.2 決定木の活用

退学の情報を目的変数として，決定木及びその拡張手法でアルランダムフォレストを用いて，説明変数を特徴量として寄与度を調べた．解析パッケージとして R 言語を用いた.説明変数の候補変数としては，表 1 のデータセットを用いた．

図 1 に決定木の結果を示す．図 1 から，授業ポータルシステムで取得した授業の出席状況と授業外学修時間で活用している e ラーニングシステムの課題達成率が高い寄与を示すことが分かった．これらをアンサンブル平均処理として評価するため，ランダムフォレストを用いて特徴量分析を行った.解析パッケージは同じく R を用いた．結果を図 2 に示す．この結果からも，出席率と e ラーニングの達成率が特徴量として高い寄与を持つことが確認できた．

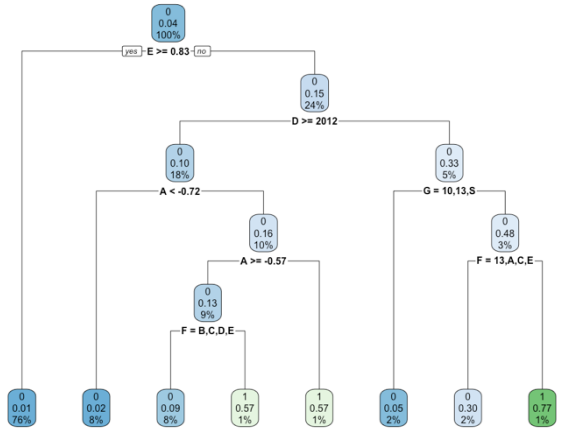

図 1 決定木の結果例

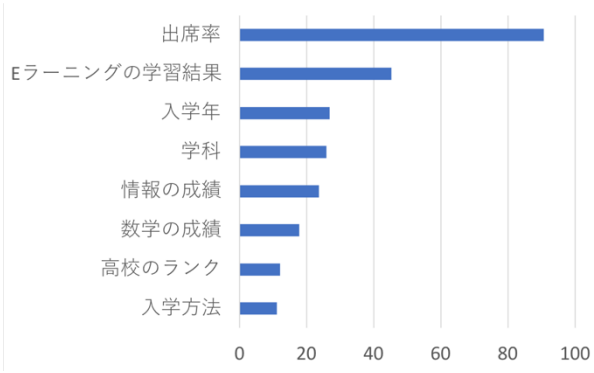

図 2 特徴量の寄与度

5. 今後

本研究では，図 2 に示した特徴量を取得できた．一方で、我々の先行研究[3]では，SOM とクラスタリングを用いた分析を通じて，高校の評定平均が特徴量として強く寄与することを示している．今後，本研究で得られた特徴量及び先行研究の結果を入力データの一部として DL に適用することで，別の特徴量の存在を調べる．また，DL の推論性能評価を通じて，適切なネットワーク構造（適した特徴量抽出可能な構造）を洗い出していく．これらをデータベース化して，退学者分析システムとしての活用の可能性を検証していく．

参考文献

(1) 文部科学省:”学生の中途退学や休学等の状況について”
(http://www.mext.go.jp/b_menu/houdou/26/10/__icsFiles/afieldfile/2014/10/08/1352425_01.pdf)(2018 年 6 月 10 日アクセス)
(2) 高橋 大樹,小松川 浩,” DL を用いた中途退学者推論に関する一検討”,2017 年度 JSiSE 学生研究発表会,(2018)
(3) 髙橋 駿嗣, 小松川 浩,” 教学 IR 支援に向けた SOM による退学者の傾向分析”,情報システム情報学会第 41 回全国大会 pp.265-266,(2016)