ICT 上の学生データを用いた中途退学者推論プログラムの改良

Improvement of inference program for dropouts 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, A 大学内の ICT 上の学生データを分析し, 在学生の中から将来的に中途退学をする学生を推論するプログラムの開発を行なっている. 本研究では, その中途退学者推論プログラムを改良し, 学生データの統計的処理を行い, 中途退学者推論の推論精度の向上を目指した.
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た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)で行った中途退学者の分析手法を用いて, 中途退学者推論プログラムの改良を行なった.

2. 本研究の目的

近年, ICT 教育支援システムが普及し, ICT 上の学生の学習データが蓄積されている. また DL の教育分野への適応が期待されている. そのため本研究では, 蓄積された学生の学習データから学生の中途退学傾向を DL を用いて解析した.

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た, 中途退学者の推論精度のさらなる向上を目的としてデータの統計的処理を行った. 統計的処理を行なった A 大学のデータを学習データとして入力し, 学生データの特徴量の分析を行った. また分析を行なった結果からデータの統計的処理を行うことで推論精度が 76%から 96%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. 先行研究で利用したデータは A 大学の複数のデータから取得し, データセットにした. データセットとは先行研究で用いるために整えたデータ群である. 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. 先行研究では, 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 列のデータを利用した. 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.

3.3 先行研究の分析

先行研究で利用したデータの中で, 欠損率が低いデータを抽出し, そのデータを利用してデータの特徴量の分析を行なった.抽出したデータは,E ラーニングの学習の進捗率, 入学年, 学科, 出席率, 高校のランク, 入学方法のデータを利用した. その際, 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.その結果学生の中途退学傾向を推論するために最も重要な特徴量は出席率であった. 次に, 出席率のデータの欠損値を補完した. その後, アンダーサンプリング, オーバーサンプリングをすることで, データを均等データに近づけた. その結果,中途退学者推論の推論精度が 96%程度となった.

3.4 先行研究の問題点

先行研究の問題点として，中途退学者推論の推論精度を判定するデータに対しても統計的処理をしてしまったため，実際のデータと異なる結果になってしまっており，推論精度が高くなってしまっていた．そのため本研究では，データに対して適切な処理を行うようにプログラムを改良することにした．

4. データの統計的処理

4.1 PMM

PMM(Predictive Mean Matching)とは多重代入法の一種である．PMM では欠損値を保管したデータセットを複数作成し，それぞれのデータセットに対して解析を行う手法である．

4.2 アンダーサンプリング

アンダーサンプリングとは不均衡データに対して標本数が多いデータをランダムに抽出する方法である．これにより不均衡データを均衡データに近づけることができる．

4.3 オーバーサンプリング

オーバーサンプリングとは，不均衡データに対して標本数が少ないデータを作成し，均衡データに近づける手法である．データの作成方法は標本数が少ないデータからその近傍にあるデータを生成し，その生成したデータとの間でランダムにデータを作成することで標本数が少ないデータを増加させることができる．

5. プログラムの改良

5.1 統計的処理

本研究ではこれまで行なっていた中途退学者推論プログラムに改良を行った．これまではデータベースから取得したデータをプログラムを用いて統計的処理を行なってから中途退学者推論プログラムに入力していた．しかし中途退学者推論プログラム自体に統計的処理を行うプログラムを入れることで，データセットの変更が容易に行うことができ，推論精度の確認が容易に行えるようになった．また先行研究の問題点を解決することができた．

5.2 交差検証

先行研究では，データセットをランダムに抽出しテストデータを作成し，それ以外のデータを学習データとして中途退学者推論プログラムに入力をしていた．しかし今回のプログラムでは交差検証を行うことでその推論精度の妥当性を検証することができるように改良した．交差検証とはデータセットをいくつかに分割し，その分割した一つを除き学習データとして学習し，残りをテストデータとして推論精度の検証を行う方法である．図 1 に交差検証の例を示す．図 1 ではデータセットを 3 分割にして，2 つを学習データ，1 つをテストデータとして学習し、検証を行なっている．学習と検証は分割した数と同等の回数繰り返す．

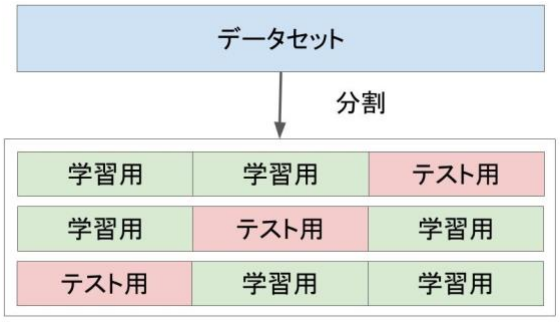

図 1 交差検証

6. 今後の取り組み

本研究では，中途退学者推論プログラムの改良を行なった．しかし学生の特徴量として中途退学の傾向だけではなく，学科の配属などの傾向を推論することができると考える．そのため今後は中途退学者推論だけではなく，学生の特徴の分析を目指す．

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)(2019 年 2 月 10 日アクセス)

(2) 高橋 大樹,小松川 浩,” 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”,第 43 回教育システム情報学会 全国大会,(2018)