

# Deep Learning を用いた中途退学者推論プログラムの構築

## Construction of inference program for dropouts using Deep Learning

高橋 大樹<sup>\*1</sup>, 小松川 浩<sup>\*1</sup>

Hiroki TAKAHASHI<sup>\*1</sup>, Hiroshi KOMATSUGAWA<sup>\*1</sup>

<sup>\*1</sup>千歳科学技術大学総合光科学部

<sup>1</sup>Faculty 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、30%程度の推論精度を得た。本研究では、入力データの増量、および学習方法の見直しを行い、推論精度の向上を図った。推論精度の評価は、3514 件の標本から 2811 件を学習させた際の、残りの 703 件を評価用のデータとして、機械学習の評価指標に基づいて行った。その結果、推論精度が 76%程度となった。

キーワード：Deep Learning 中途退学者推論 データ プログラム

### 1. はじめに

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。特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<sup>(1)</sup>。

そこで本研究では、先行研究の中途退学者推論プログラム<sup>(2)</sup>を用いて、入力データの増量、および学習方法の見直しを行い、推論精度の向上を図った。推論精度の評価は、3514 件の標本から学習比率を 80:20 とし機械学習の評価指標に基づいて行なった。

### 2. 中途退学者の推論方法

本研究では先行研究と同様に A 大学での中途退学者を含むこれまでの学生のデータを用いて、退学者の推論を行った。手順としては、まず、学生に関する情報を抜き出し、DL が読み込めるような形にデータを整えた。その後、中途退学者の傾向にある学生を推論するプログラムを用いて層構造や学習回数を変更することで推論精度がどのように変化していくかを検証した。また計算量が増加することが見込まれるため GPU を利用して高速化を行った。

#### 2.1 データセット

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 大学が研究開

発・運用を行っている 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から取得している。それぞれ三つのデータベースには学生の退学者推論とは関係がないと思われる教員の個人情報、父母の個人情報などのデータも含まれていたため、それらを除いたデータをデータセットとした。データセットとは本研究で用いる中途退学者推論へ入力できるよう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合計で 444 名である。先行研究ではデータが出席率、GPA、演習時間、テスト結果、評定平均、高校ランク、在籍状況の 7 データであったが、本研究では、先行研究のデータに加えて学内ポータルサイト、学習カルテ、e ラーニングシステムの 3 つのデータベースから得た 30 のデータを使用した。合計したカラム数は計 37 データである。

#### 2.2 中途退学者推論の学習回数と中間層の関係

中途退学者推論のための学習回数の初期値は 100 回に設定し、増分 100 回で 1000 回までのデータを取得した。中間層の構造は最初 3 層に設定しノード数を  $30 > 20 > 10$  と減らしていった。次に層を 4 層に設定しノード数を  $30 > 25 > 20 > 10$  と減らしていった。また、次も同様に 5 層にしてノード数を  $30 > 25 > 20 > 15 > 10$  と減らしていった。またノード数を増やす構造も検証した。具体的には 3 層のときはノード数を  $10 > 20 > 30$  に設定し、4 層のときはノード数を  $10 > 15 > 20 > 30$  に設定し、5 層のときはノード数を  $10 > 15 > 20 > 25 > 30$  に設定した。実際に DL で学習を行う際に、データを学習させる単位を `batchsize` と定義した。学習回数と中間層の構造の関係を検証するときは `batchsize` を 595 で固定とした。学習及び推論は 500 回繰り返す、推論精度を分析し、平均を求め

た.

### 2.3 batchsize と学習回数との関係

中間層数やノード数, batchsize, 学習回数を変更し推論精度の向上を検証した結果, batchsize が小さいほど, 推論精度が増加する傾向にあることが判明した. そして, 学習回数が増加すると batchsize の最小値が下がるという関係性が判明した. これは学習回数を増やすことで一度に学習することが出来るデータが少なくても推論が可能なたためである. したがって,

### 3. 中途退学者の推論結果

精度の向上を図るため, 学習比率を学習に扱うデータを 8 割, テストに用いるデータを 2 割に設定し, 学習回数を 1000 回から増分 1000 回で, 10000 回まで検証した. 1000 回毎に batchsize をエラーが起きない最小値とし, 推論を行った. 中間層は 5 層とし, ノード数が 10 > 15 > 20 > 25 > 30 と増加していく設定に固定した. プログラムは 100 回繰り返し, 推論精度を分析し, 平均を求めた. 分析結果から, 推論精度が約 76%から 74%の間で変化することがわかった. 結果を図 1 に示す. 図 1 から, 学習回数が 1000 回のは推論精度が最も高く, その後は推論精度が下がり続け, 学習回数が 7000 回ときに一度だけ推論精度が少し上がるがすぐに推論精度が下がっていることがわかる. batchsize の最小値は学習回数が増えるとともに下げることが出来るが, 546 が限界であった. 学習回数が 1000 回から下がり続けていることから, 学習回数 1000 回が最適であると考えられる.

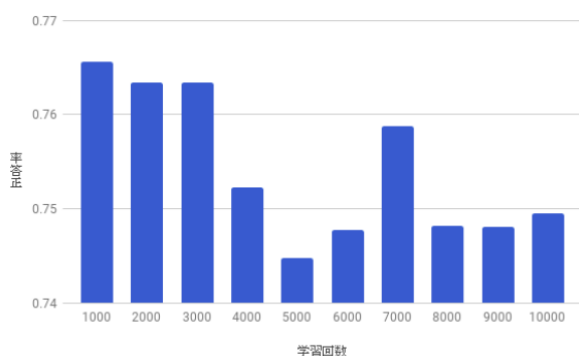


図 1 batchsize を最小にしたときの学習回数による推論精度の変化

### 4. まとめ

本研究では先行研究の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. 学習データは, 3514 名のうち 444 名が退学者として記録されているデータを扱った. 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>15>20>25>30 と増加していく構造, 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%程度であることがわかった. またデータには情報が入力されていない部分が有ったため, 情報が完全に入ったデータを扱うことで精度向上につながる可能性がある.

### 5. 今後の取り組み

本研究で推論することが出来た推論精度の最も高い数値が 76%程度であった. しかし, 76%の数値ではまだ実用的とは言えない. さらに精度を上げる手法として, 本研究で扱った学生の情報はデータが欠損している部分が多かったため, できるだけデータが埋まっている学生の情報を採用することで精度を上げる必要がある. また, 本研究ではデータを整理して選別を行った結果 37 のデータが残ったが, 170 近くあったデータから使えるものを選別したため手作業では限界が生じている. そのため, 選別する作業を自動化し, データセットを作成してくれるシステムを開発することが今後の課題として挙げられる.

#### 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/afildfile/2014/10/08/1352425\\_01.pdf](http://www.mext.go.jp/b_menu/houdou/26/10/_icsFiles/afildfile/2014/10/08/1352425_01.pdf)) (2017 年 12 月 27 日アクセス)
- (2) 高橋 良貴, 小松川 浩, ”DL を用いた中途退学者推論に関する一検討”, 2016 年度 JSiSE 学生研究発表会, 2017 年 3 月 4 日