オープンデータを活用した学習分析研究者のための ディープラーニングセミナー
ー学習分析学会における取り組みー

Deep Learning Seminar for Learning Analytics Researchers using Open Data
- An Initiative by Japanese Society for Learning Analytics -

卯木 輝彦[*1,*6], 加藤 利康[*2], 佐藤 伸也[*3], 堤 宇一[*4,*6], 児玉 靖司[*5,*6]
Teruhiko UNOKI[*1,*6], Toshiyasu KATO[*2], Shinya SATO[*3], Uichi TSUTSUMI[*4,*6], Yasushi KODAMA[*5,*6]
[*1] 株式会社フォトロン, [*2] 日本工業大学, [*3] 茨城大学, [*4] アルー株式会社, [*5] 法政大学
[*1]Photron Limited, [*2]Nippon Institute of Technology, [*3]Ibaraki University, [*4]Alue Co., Ltd., [*5]Hosei University
[*6] 学習分析学会
[*6]Japanese Society for Learning Analytics
Email: unoki@photron.co.jp

あらまし：学習分析学会では，学習分析へのディープラーニングの適用拡大を目的に，2016 年から「ディープラーニングによる Learning Analytics ワークショップ」を開催してきた．その後，ディープラーニングの応用を体系的に学びたいとの要望を受け，ハンズオン形式のセミナーを実施することにした．セミナーでは実践的な学習ができるよう，公開されている実データを使い実習を行っている．本稿では，JASLA 主催ディープラーニングセミナーの概要を紹介し，セミナーで使用したデータセットについて述べる．
キーワード：ディープラーニング，学習分析，ラーニングアナリティクス，オープンデータ，セミナー

1. はじめに

教育データにディープラーニングを適用した研究は，2015 年頃からみられる(1)．初期の研究は，オンラインコースや LMS の操作ログなど，受講生の PC 操作履歴データを利用して，成績予測や離脱予測を行う研究が多い．その後，グループワークやプレゼンテーション練習など，学習者の身体の動きや心的状態が重要となる学習活動を対象に，映像，音声，その他各種センサーデバイスから得られたデータに対してディープラーニングを適用する研究がみられるようになった(2)．

学習分析学会（JASLA）では，Learning Analytics（LA）研究の普及促進を目的に各種研究会やセミナー等のイベントを実施している．2017 年 2 月および 2018 年 3 月の二度にわたり「ディープラーニングによる LA ワークショップ」（以下，LA ハッカソン）を開催した(3)．

LA ハッカソンでは，参加者は数名ずつのグループに分かれ，成績予測やクラスタリングなどの課題に取り組んだ．大学で実際に運用中の e ラーニングシステムから収集した学習履歴データを使用した．ディープラーニングを実際の教育データに適用することで，従来手法との差異や適用するための課題を把握するとともに，使用したデータセットにおける学習行動と成績の関係を明らかにした．

一方，LA ハッカソンで使用したデータセットは，匿名化を施してはいるが，成績などの情報が含まれている．データを提供いただいた大学からの使用許可は，プライバシー保護の観点から本イベントに限定されていた．すなわち，参加者は LA ハッカソンで使用したデータを持ち帰って振り返り学習等に活用することができなかった．

LA ハッカソンの限られた時間内だけでディープラーニングの多くを習得することは容易ではない．参加者の多くから，自由に使えるデータを使って，基礎的な事項から体系的にディープラーニングを学びたいとの要望があがってきた．そこで，JASLA では，ディープラーニングの LA への応用に関心を持つ方を対象に，ハンズオン形式のセミナーを実施することにした．

本稿では，JASLA ディープラーニングセミナーの概要を紹介し，セミナー内で使用したデータセットについて述べる．

2. JASLA ディープラーニングセミナー

セミナーは，2017 年 12 月から 2019 年 2 月までに計 4 回を実施した(4)．参加者は，毎回満席で 25 名程度，内訳は大学教員が 6 割，学生が 3 割，残りが企業である．ディープラーニングの可能性は知りつつも，これまで手を動かして学ぶ機会を作れなかった方や，環境構築でつまずいていた方が多い．参加者にはノート PC を持参してもらい，各自が自分の PC でプログラムの実行結果を確認できるようにした．

実行環境は，セミナーの前半 2 回と後半の 2 回で大きく変えた．前半 2 回では各自の PC に Python 環境を構築した．当時，クラウド上で容易に機械学習が実行できる適当な環境が存在しなかったためである．ノート PC の性能を考慮すると，大きなサイズのデータを扱うことはできなかった．後半 2 回は状況が変わり，Google Colaboratory(5)を使うことにした．Google Colaboratory は，クラウド上で実行される Python の実行環境である．Google アカウントさえあれば，GPU/TPU も無料で利用できる．Google Drive や Github との連携ができるため，参加者へのデータ

の配布も容易で, PC の性能によらず比較的大きなデータを扱うことが可能になった.

セミナーでは，Python や機械学習の基礎的な学習項目に加え，毎回異なる題材を取り上げ，実データを使い様々なディープラーニング手法での分析を行った．題材として，前半 2 回は LMS のデータを利用した成績予測などを，後半 2 回は時系列データを利用した異常検知や行動分析などを取り上げた.

3. データセット

3.1 利用するデータセットの選定

セミナーで使用するデータセットは，次の条件を満たすものとした.

(1) インターネット経由でだれもが入手可能
(2) 他の研究者による分析結果が公開されている
(3) セミナー環境で妥当な時間内に処理が可能

データセットの選定は，各所で運用されている機械学習用データリポジトリ，Kaggle[6]に代表される機械学習コンペティションプラットフォーム，edX や Coursera などのオンラインコース，関連学会で開催された機械学習コンペティションやチュートリアルなどから情報を得て行った.

例えば，KDD Cup 2015[7]では，中国の MOOC 受講者の離脱確率予測を題材としたデータセットが利用できるが，比較的データのサイズが大きく複雑な構造のため，セミナーでの利用は見送った．PSLC DataShop[8]では，知的学習支援システムで収集された学習者の行動履歴が蓄積されているが，結果の比較ができそうな先行研究が少なく, 利用を見送った.

時系列データについては，実際の教育の場で取得された適当なオープンデータを発見することができなかった．今後のマルチモーダル LA 研究でも有用と考え得る波形データを中心に，直観的に理解しやすいデータを採用することにした.

以下では，セミナーで使用したデータセットの一部について概観する.

3.2 Students' Academic Performance Dataset[9]

ヨルダン大学の学習管理システム（LMS）によって収集された 480 名分の学生のデータから成る．e ラーニングシステム上のグループディスカッションへの参加回数や欠席状況など計 16 の特徴量を持ち，各個人の最終成績が 3 段階で付与されている.

3.3 Classifying Heart Sounds Challenge

心音データを分類するデータ分析コンペティション Classifying Heart Sounds Challenge 2011[10]で使用されたデータセットである． 30 秒以下で心音を録音した長さの異なる 812 個の音声ファイルから成る.

3.4 MIT-BIH Arrhythmia Database[11]

男女 47 人の被験者から得られた 48 時間 30 分の心電図データである. 心臓専門医が付与した約 11 万個のアノテーションデータとともに公開されている.

3.5 Human Activity Recognition Data Set[12]

19 歳から 48 歳の 30 人が，スマートフォンを腰に着用し，6 種類の活動（歩行，階段上昇，階段下降，着席，直立，横たわる)を行い収集されたデータである．スマートフォンで計測した角加速度，直線加速度およびジャイロのデータからなる.

4. まとめ

JASLA ディープラーニングセミナーについて，セミナーで使用したデータセットを中心に述べた.

画像認識の分野では研究や教育に利用できる大規模なデータセットが多数公開されている．実験結果を実行可能なプログラムとともに公開することも盛んである．学習分析の分野においても学習履歴データを共有しようとする動きは広がりつつあるが，現時点ではその数は少ない．オープンなデータセットが数多く公開されることを期待したい.

参考文献

(1) Coelho, O. B. and Silveira, I. :“Deep Learning applied to Learning Analytics and Educational Data Mining: A Systematic Literature Review”, Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE), pp.143-152 (2017)
(2) Baker, R. et al.:“Workshop on deep learning with educational data”, Proceedings of the 10th International Conference on Educational Data Mining, p.474 (2017)
(3) 児玉靖司，卯木輝彦，加藤利康，堤宇一：“学習分析学会研究活動報告”，学習分析学会 2018 年度第 1 回研究会 (2018)
(4) 学習分析学会：“Keras/Python3 で学ぶ ディープラーニングによる時系列データ解析入門【実践編】”, https://jasla.jp/event/seminar007/ (2019 年 6 月 14 日アクセス)
(5) Google Colaboratory，https://colab.research.google.com/ (2019 年 6 月 14 日アクセス)
(6) kaggle, https://www.kaggle.com/ (2019 年 6 月 14 日アクセス)
(7) KDD Cup 2015, https://www.kdd.org/kdd2015 (2019 年 6 月 14 日アクセス)
(8) PSLC DataShop , https://pslcdatashop.web.cmu.edu/ (2019 年 6 月 14 日アクセス)
(9) kaggle: "Students' Academic Performance Dataset", https://www.kaggle.com/aljarah/xAPI-Edu-Data (2019 年 6 月 14 日アクセス)
(10) Bentley, P. et al.:“The PASCAL Classifying Heart Sounds Challenge 2011 Results ”, http://www.peterjbentley.com/heartchallenge/, (2019 年 6 月 14 日アクセス)
(11) Goldberger, A.L. et al.:“PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals”, Circulation 101(23) , pp.215-220 (2000)
(12) UCI Machine Learning Repository:“Human Activity Recognition Using Smartphones Data Set ”, https://archive.ics.uci.edu/ml/datasets/human+activity+recognition+using+smartphones (2019 年 6 月 14 日アクセス)