Assessing the Utility of Deep Learning: Using Learner-System Interaction Data from BioWorld

Tenzin Doleck
University of Southern California, USA
doleck@usc.edu

Eric G. Poitras
University of Utah, USA
eric.poitras@utah.edu

Susanne P. Lajoie
McGill University, Canada
susanne.lajoie@mcgill.ca

Abstract: In recent years, deep learning (LeCun, Bengio, & Hinton, 2015) has drawn interest in many fields. As optimism for deep learning grows, a better understanding of the efficacy of deep learning is imperative, especially in analyzing and making sense of educational data. This study addresses this issue by establishing a benchmark for a common prediction task – student proficiency in diagnosing patient diseases in a system called BioWorld (Lajoie, 2009). To do so, we compared deep learning to existing solutions, including traditional machine learning algorithms that are commonly used in educational data mining. The dataset consists of log interaction data collected from 30 medical students solving 3 different cases. A 10-fold cross-validation method was used to evaluate the predictive accuracy of each model. Interestingly, our results indicate that deep learning does not outperform traditional machine learning algorithms in predicting diagnosis correctness. We discuss the implications in terms of understanding the proper conditions for its use in educational research.

Keywords: Machine learning; deep learning; educational data mining; computer-based learning environments; medical education

Introduction

Machine learning algorithms have been the focus of considerable research, resulting in significant progress made in various disciplines. Among the plethora of algorithms, perhaps, one of the most promising is deep learning, which "allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction" (LeCun, Bengio, & Hinton, 2015, p. 436). Several recent reviews have generally addressed the benefits and opportunities of deep learning (Deng & Yu, 2014; Guo, Liu, Oerlemans, Lao, Wu, & Lew, 2016; Miotto, Wang, Wang, Jiang, & Dudley, 2017). Indeed, deep learning has been successfully applied in various contexts and settings (LeCun et al., 2015; Zhang, Yang, Chen, & Li, 2018), ranging from finance (Heaton, Polson & Witte, 2016) to medical image analysis (Litjens et al., 2017). Yet it has only recently been gaining the attention of the educational data mining community (Wilson et al., 2016), and its application to educational data has yet to be fully realized (Xiong, Zhao, Van Inwegen, & Beck, 2016).

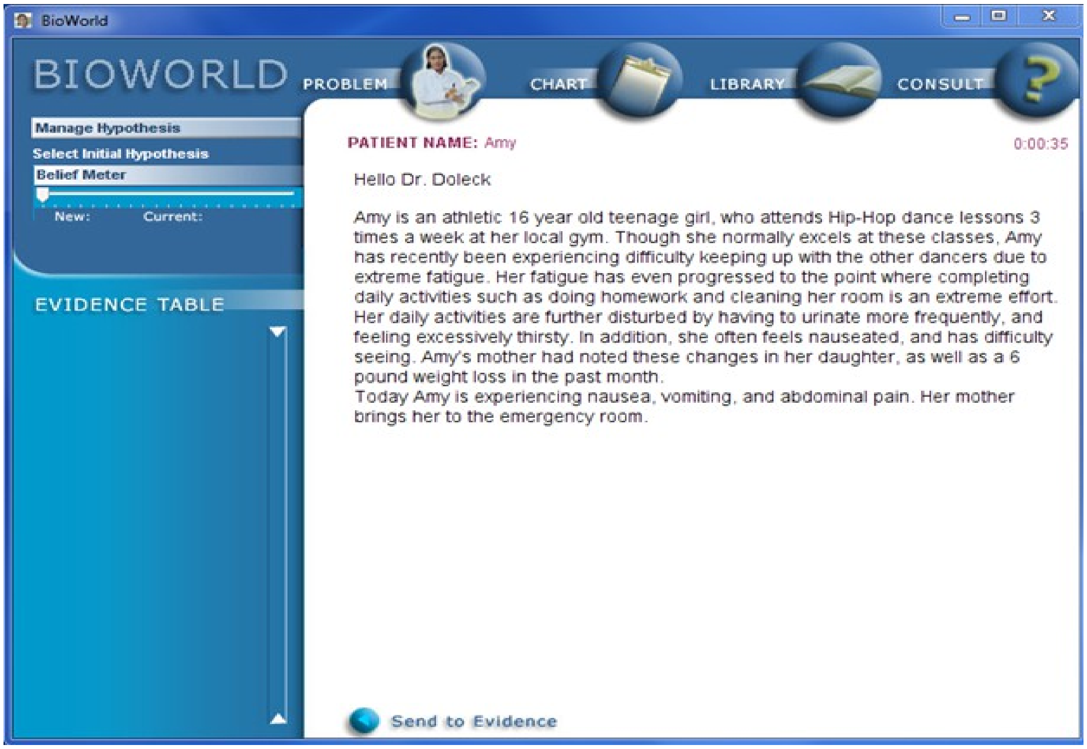
As learners interact with and use computer-based learning environments, these systems often log their actions. The data thus collected has, over the past decade, led to the considerable growth of research related to educational data mining (EDM) (Baker & Inventado, 2014; Bakhshinategh, Zaiane, ElAtia, & Ipperciel, 2018;

Papamitsiou & Economides, 2014). Research in this interdisciplinary subfield has especially emphasized the importance of machine learning techniques to acquire insights into learning (Doleck, Basnet, Poitras, & Lajoie, 2015; Dutt, Ismail, & Herawan, 2017; Peña-Ayala, 2014; Poitras, Doleck, & Lajoie, 2018; Romero & Ventura, 2012). Despite the significant progress made in this area, relatively few studies have applied deep learning algorithms to examine educational data such as clickstream or log transaction data (Botelho, Baker, & Heffernan, 2017; Jiang et al., 2018).

Motivated by the general inattention to and the need to better understand and assess the utility and performance of deep learning in varied educational contexts, the current study explores the use of deep learning. Specifically, we apply a deep learning package (WekaDeeplearning4J) on learner-system interaction data from a medical computer-based learning environment called BioWorld. The task to which we apply deep learning relates to predicting clinical reasoning performance of medical students (i.e., diagnosis correctness), a common task in educational data mining to establish student performance profiles. The details regarding the data used in the current study are provided in the following section.

Context: BioWorld

BioWorld (Figure 1) is a computer-based learning environment designed to help medical students regulate their learning about medical reasoning with virtual patients while receiving feedback (Doleck, Jarrell, Poitras, Chaouachi, & Lajoie, 2016; Poitras, Lajoie, Doleck, & Jarrell, 2016). The BioWorld system captures user interactions while medical students solve virtual patient cases (Lajoie, Poitras, Doleck, & Jarrell, 2015), such as: attempt identifier (participant and case ID), a timestamp, the BioWorld space (e.g., chart), the action taken (e.g., add test), and details in relation to the action (e.g., Cortisol Result: 138-690 nmol/L (0800 hr); 138-414 nmol/L (1600 hr)). Analyzing such data can provide insights into the use of the learning material and facilitate a deeper understanding of the learning process and outcomes.



PATIENT NAME: Amy

0:00:35

Hello Dr. Doleck

Amy is an athletic 16 year old teenage girl, who attends Hip-Hop dance lessons 3 times a week at her local gym. Though she normally excels at these classes, Amy has recently been experiencing difficulty keeping up with the other dancers due to extreme fatigue. Her fatigue has even progressed to the point where completing daily activities such as doing homework and cleaning her room is an extreme effort. Her daily activities are further disturbed by having to urinate more frequently, and feeling excessively thirsty. In addition, she often feels nauseated, and has difficulty seeing. Amy's mother had noted these changes in her daughter, as well as a 6 pound weight loss in the past month.
Today Amy is experiencing nausea, vomiting, and abdominal pain. Her mother brings her to the emergency room.

Figure 1. Bioworld Interface

Data

The data file used in the current study contains aggregate student interaction data collected from over thirty participants (19 women and 11 men, with an average age of 23 (*SD*=2.60)) and three endocrinology cases: diabetes mellitus (type 1), hyperthyroidism, and, pheochromocytoma. For the predictive modeling task, the following features were extracted from the log trace data to characterize student performance: problem type; number of lab tests ordered; count of evidence items and lab tests ordered similar to expert solution; time taken to solve the case; and confidence in the diagnosis.

Analyses and Results

All analyses were conducted in the WEKA workbench, which includes various machine learning algorithms (Hall, Frank, Holmes, Pfahringer, Reutemann, & Witten, 2009). A set of commonly used classification algorithms and a deep learning algorithm were tested on the dataset to establish a benchmark in predictive accuracy, which include the following: NaiveBayes, Logistic, SMO, IBK, JRip, J48, and WekaDeeplearning4J. The models were trained and validated with 10-fold cross-validation.

We present the results of the 10-fold cross-validation evaluation for the different classifiers in Table 1. As illustrated in Table 1, performance was observed to be above chance (Kappa>0 and AUC>0.5) (Hulse, Harrison, Ostrow, Botelho, & Heffernan, 2018; Jiang et al., 2018). The classification accuracies ranged from 73.33% to 84.44%. In terms of predictive performance, our results reveal that the deep learning package, WekaDeeplearning4J, does not outperform other commonly used algorithms (NaiveBayes, Logistic, SMO, IBK, JRip, and J48) in the WEKA workbench.

Classifier	**Correctly Classified % (Accuracy)**	**Kappa**	**AUC**
WekaDeeplearning4J	74.44%	0.3850	0.766
NaiveBayes	76.67%	0.4119	0.765
Logistic	82.22%	0.5355	0.790
SMO	75.56%	0.2511	0.600
IBK	73.33%	0.3032	0.641
JRip	84.44%	0.6214	0.775
J48	82.22%	0.5466	0.761

Table 1. 10-fold cross-validation results

Discussion

The current research was motivated by the need to assess the applications of deep learning for educational data, which has shown promising results in a variety of domains over the last decade (see Baker, Beck, Chi, Heffernan, & Mozer, 2017). Specifically, we compared the use of deep learning to traditional machine learning algorithms in terms of existing solution to the problem of predicting diagnosis correctness in BioWorld. The findings indicate that deep learning (WekaDeeplearning4J algorithm) does not outperform traditional machine learning algorithms in predicting diagnosis correctness. Such a finding tempers optimism regarding the potentialities of deep learning and highlights the importance of considering the context to which deep learning is applied. In fact, researchers, such as Jiang et al. (2018), note inconsistencies in the reported findings. Moreover, it is not always clear how best to use deep learning. As Wilson et al. (2016) put it, “deep learning has a promising future in educational data mining, but that future depends on data sets that have a much richer encoding of the exercises and learning context” (p. 7). As such an important challenge for educational data mining research is to better understand the applicability of deep learning.

Although the current study offers insights into the applications of deep learning algorithms, there are limitations that must be recognized. It should be noted that we have examined a very specific use for the model – establishing a benchmark for predicting student proficiency at the case-level; as such, it is possible that the relative underperformance of deep learning algorithm is limited to the scope of this particular task. An alternative approach might entail predicting misconceptions or errors made during problem-solving, which may also warrant more sophisticated topologies in designing the neural network as student interactions unfold over time. The benefit of this approach to educational data may rely in gaining insights into unstructured interactions, where series of layers may be designed to make sense of data with larger amounts of dimensions that are not fully captured by traditional approaches. The generalizability of these results should also be ascertained while taking into consideration the relatively small sample size of the dataset. To conclude, better understanding the use of deep learning in educational contexts remains an important area for future work.

References

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), *Learning analytics* (pp. 61-75). New York, NY: Springer.

Baker, R., Beck, J. E., Chi, M., Heffernan, N. T., & Mozer, M. (2017). Workshop on deep learning with educational data. In *Proceedings of the 10th International Conference on Educational Data Mining Workshops and Tutorials*.

Bakhshinategh, B., Zaiane, O., ElAtia, S., & Ipperciel, D. (2018). Educational data mining applications and tasks: A survey of the last 10 years. *Education and Information Technologies, 23*(1), 537-553. doi: 10.1007/s10639-017-9616-z

Botelho, A. F., Baker, R. S., & Heffernan, N. T. (2017). Improving sensor-free affect detection using deep learning. In E. André, R. Baker, X. Hu, Ma. M. T. Rodrigo, B. Boulay (Eds.), *Proceedings of the 18th International Conference on Artificial Intelligence in Education* (pp. 40-51). London, UK: Springer International Publishing.

Deng, L. & Yu, D. (2014). Deep Learning: Methods and Applications. *Foundations and Trends in Signal Processing, 7*(3-4), 197-387. doi: 10.1561/2000000039

Doleck, T., Basnet, R. B., Poitras, E. G., & Lajoie, S. P. (2015). Mining Learner-System Interaction Data: Implications for Modeling Learner Behaviors and Improving Overlay Models. *Journal of Computers in Education, 2*(4), 421–447. doi:10.1007/s40692-015-0040-3

Doleck, T., Jarrell, A., Poitras, E. G., Chaouachi, M., & Lajoie, S. P. (2016). A Tale of Three Cases: Examining Accuracy, Efficiency, and Process Differences in Diagnosing Virtual Patient Cases. *Australasian Journal of Educational Technology, 36*(5), 61–76. doi:10.14742/ajet.2759

Dutt, A., Ismail, M., & Herawan, T. (2017). A Systematic Review on Educational Data Mining. *IEEE Access, 5*, 15991-16005. doi: 10.1109/access.2017.2654247

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., & Lew, M. (2016). Deep learning for visual understanding: A review. *Neurocomputing, 187*, 27-48. doi: 10.1016/j.neucom.2015.09.116

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The WEKA data mining software. *ACM SIGKDD Explorations Newsletter, 11*(1), 10. doi: 10.1145/1656274.1656278

Heaton, J., Polson, N., & Witte, J. (2016). Deep learning for finance: deep portfolios. *Applied Stochastic Models in Business and Industry, 33*(1), 3-12. doi: 10.1002/asmb.2209

Hulse, T., Harrison, A., Ostrow, K. S., Botelho, A. F., & Heffernan, N. T. (2018). Starters and Finishers: Predicting Next Assignment Completion from Student Behavior During Math Problem Solving. In K. E. Boyer & M. Yudelson (Eds.), *Proceedings of the 11th International Conference on Educational Data Mining* (pp. 525-528). Buffalo, NY: International Educational Data Mining Society.

Jiang, Y., Bosch, N., Baker, R., Paquette, L., Ocumpaugh, J., Andres, J. M. A. L., Moore, A.L., Biswas, G. (2018). Expert Feature-Engineering vs. Deep Neural Networks: Which is Better for Sensor-Free Affect Detection? In C. P. Rosé, R. Martínez-Maldonado, U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. McLaren, & B. du Boulay (Eds.), *Proceedings of the 19th International Conference on Artificial Intelligence in Education* (pp. 198-211). London, UK: Springer International Publishing.

Lajoie, S. P. (2009). Developing professional expertise with a cognitive apprenticeship model: Examples from Avionics and Medicine. In K. A. Ericsson (Ed.), *Development of professional expertise: Toward measurement of expert performance and design of optimal learning environments* (pp. 61–83). Cambridge: Cambridge University Press.

Lajoie, S. P., Poitras, E. G., Doleck, T., & Jarrell, A. (2015). Modeling metacognitive activities in medical problem-solving with bioworld. In A. Peña-Ayala (Ed.), *Metacognition: Fundaments, applications, and trends* (pp. 323–343). New York: Springer Series: Intelligent Systems Reference Library.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436-444. doi: 10.1038/nature14539

Litjens, G., Kooi, T., Bejnordi, B., Setio, A., Ciompi, F., & Ghafoorian, M. et al. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis, 42*, 60-88. doi: 10.1016/j.media.2017.07.005

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. (2017). Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics, 19*(6), 1236-1246. doi: 10.1093/bib/bbx044

Papamitsiou, Z., & Economides, A. (2014). Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence. *Educational Technology & Society, 17*(4),49–64

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications, 41*(4), 1432-1462. doi: 10.1016/j.eswa.2013.08.042

Poitras, E. G., Lajoie, S. P., Doleck, T., & Jarrell, A. (2016). Subgroup Discovery with User Interaction Data: An Empirically Guided Approach to Improving Intelligent Tutoring Systems. *Educational Technology & Society, 19*(2), 204–214.

Poitras, E., Doleck, T., & Lajoie, S. (2017). Towards detection of learner misconceptions in a medical learning environment: a subgroup discovery approach. *Educational Technology Research and Development*, *66*(1), 129-145. doi: 10.1007/s11423-017-9555-9

Romero, C., & Ventura, S. (2012). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3*(1), 12-27. doi: 10.1002/widm.1075

Wilson, K. H., Xiong, X., Khajah, M., Lindsey, R. V., Zhao, S., Karklin, Y., ... & Heffernan, N. (2016). Estimating student proficiency: Deep learning is not the panacea. In *Neural Information Processing Systems, Workshop on Machine Learning for Education*.

Xiong, X., Zhao, S., Van Inwegen, E. G., & Beck, J. E. (2016). Going deeper with deep knowledge tracing. In *Proceedings of the 9th international conference on Educational Data Mining* (EDM 2016) (pp. 545–550).

Zhang, Q., Yang, L., Chen, Z., & Li, P. (2018). A survey on deep learning for big data. *Information Fusion, 42*, 146-157. doi: 10.1016/j.inffus.2017.10.006