A DEEP LEARNING ANALYTICS TO FACILITATE SUSTAINABILITY OF STATISTICS EDUCATION

Taerim Lee
Korea National Open University, Rep. of Korea
trlee@knou.ac.kr

Deep Learning Analytics uses predictive models that provide actionable information. It is a multidisciplinary approach based on data processing, AI technology-learning enhancement, educational data mining, and visualization. The problem is that embracing DLA(Deep Learning Analytics) in evaluating data in higher education diverts educators' attention from clearly identifying methods, benefits, and challenges of using DLA in higher education. Predictive models including random forest (RF), support vector machines (SVM), logistic regression (logistic), and Deep Learning were trained and their performances compared. The predicted value of "source of sustainability" and selected input variables were utilized to predict the drop out of learner. Expected significant outcomes and impact is that using DLA we can find the optimal learning management model for supporting services for instructors significantly impact the quality of statistics education and for learners is necessary to support announcements from instructors, for providing appropriate learning environments.

INTRODUCTION AND BACKGROUND

Data analysis provides educational stakeholders a comprehensive overview of the performance of the institution, curriculum, instructors, students, and post-educational employment outlooks. It also provides scholars and researchers with needed information to identify gaps between education and industry so that educators and institutions can overcome these deficiencies in course offerings. More important, the ability of big data to provide these revelations can help the field of education make significant progress to improve learning processes.

Deep learning analytics uses predictive models that provide actionable information. It is a multidisciplinary approach based on data processing, AI technology-learning enhancement, educational data mining, and visualization (Scheffel, Drachsler, Stoyanov, & Specht, 2014). AI is in its embryonic stage yet companies are being formed by leveraging current AI models in education. The report forecasts the global AI in education market to grow from USD 537.3 million in 2018 to USD 3.6 billion by 2023, at a Compound Annual Growth Rate of 47.0% during the forecast period (Business Wire, 2018).

In the 10th International Conference on Educational Data Mining there was a workshop on deep learning with educational data where focuses on application of deep learning for educational data. There were variety topics of deep learning: new prediction and modeling problems, best practices for featuring data, network architectures, approaches to pre-training and interpreting learned models, end-to -end deep learning approaches with low level non-symbolic data, toolkits developed, empirical results.

The purpose of DLA is to tailor educational opportunities to the individual learner's need and ability through actions such as intervening with students at risk of drop out which was reported 6% of HarvardX open online course completion and 16% completion rate of KNOU graduation or providing feedback and instructional content. Conversely, educational data mining tries to generate systematic and automated responses to learners. While DLA focuses on the application of known methods and models using deep learning which is AI with Neural Network together to address alarm issues affecting student's successful learning and the organizational learning system, educational data mining focuses on the development of new computational data analysis methods.

Deep Learning Analytics algorithms extract high-level, complex abstractions as data representations through a hierarchical learning process. A key benefit of Deep Learning Analysis is the analysis and learning of massive amounts of unsupervised data, making it a valuable tool for Big Data Analytics (Najafabadi et al. 2015; Alsheikh et al. 2016; Scheffel et al. 2014; DiCerbo 2014) where raw data is largely unlabeled and un-categorized. In the present study, we explore how Deep Learning can be utilized for addressing some important problems in Big Data Analytics which come from learning process, including extracting complex patterns from massive volumes of data, semantic indexing, data tagging, fast information retrieval, and simplifying discriminative tasks. We also investigate some aspects of Deep Learning Analytics research that need further exploration to incorporate specific challenges introduced by Big Data Analytics (Najafabadi et al. 2015), including streaming data, high-

In: S Budgett (Ed.), *Decision Making Based on Data*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
August 2019, Kuala Lumpur, Malaysia. iase-web.org/Conference_Proceedings.php

dimensional data, scalability of models, and distributed computing (Vieira, Goldstein, Purzer, & Magana, 2016).

In spite of the attention given to analytics as a concept and the development of methods (statistical analysis, "big data"), educators do not have access to integrated LA system that allow for varied and complex evaluations of learner performance and comparisons between different sets of learners (Ali, Hatala, Gašević, & Jovanović, 2011). Learners also lack needed information about their performance. DLA can similarly contribute to learner motivation by providing detailed information about their performance. In response to this weakness, we propose the development of an integrated and extensible toolset that can assist academics and organizations evaluate learner activity, determine needed interventions, and improve advancement of learning opportunities. The following table shows the summary of the educational and data mining tasks addresses by the works that use Deep Learning.

This LA system will be learner facing as well, permitting individuals to track their own progress and take advantage of analytics in improving their learning activities (Business Wire, 2018). This provides for more accurate analyses and subsequent learner and instructor recommendations as well as identifying particular learning and teaching activities that promote individual student success. An understanding of the learning and teaching context for the course offerings will also assist in addressing the need for institutional comparisons and benchmarking.

RESEARCH QUESTION AND HYPOTHESIS

The problem is that embracing DLA (Deep Learning Analytics) in evaluating data in higher education diverts educators' attention from clearly identifying methods, benefits, and challenges of using DLA in higher education. These three key components need further clarification for higher education stakeholders to help them effectively apply learning analytics in higher education. Educators have to go through the daunting task of sifting through the literature to become familiar with DLA methods, benefits, and challenges. To help solve the problem, the following questions guided this review:

1. What are the methods for conducting deep learning analytics in education?
2. What are the benefits of using deep learning analytics in education?
3. What are the challenges of using learning analytics in education?

Further identifying and describing DLA methods, benefits, and challenges can better help educators in higher education to incorporate DLA to improve students' learning. DLA would be a suitable way to measure students' performances assuming big data settings. Data-driven strategies would be suggested to cope with various students' needs in such a kind of big data atmosphere for lots of purposes.

DETAILED PROJECT DESCRIPTION

Our goal was to utilize statistical models to formulate a Deep Learning Analysis system utilizing learning process information available within a few hours of learners presentation to predict the source, need for intervention and disposition in learners during open and distance learning. Using Deep Learning Analysis, it is possible to predict student performance or dropout, which makes it possible for students to identify their weakness.

There are a number of well-known machine learning (Armayor & Leonard, 2010) tools for analyzing data, including Deep learning, artificial neural networks (ANN) (Cho & Saul, 2009), k-nearest neighbor (kNN), decision trees, support vector machines (SVM). An ensemble uses the predictions of multiple base classifiers through majority voting. Boosting, a meta-classifier, combines weak classifiers and takes a weighted majority vote of their predictors. Breiman (2001) developed the random forest (RF) method by combining classification tree predictors. Ensemble methods could help improve your statistical results by combining several models. The study objective was to develop and compare the performance of machine learning models as described above to predict learner's outcomes in learners presenting with drop out.

Data collection

The purpose of data collection was to find empirical studies including quantitative, qualitative, mixed methods, and literature reviews published in peer-reviewed journals since 2000 to identify methods, challenges, and benefits of LA in higher education. Researchers collect data from the learners, process the data into metrics, and use the results to perform an intervention that affects the students. The

cycle continues as researchers collect additional data from the students for the next cycle of learning analytics. For Deep Learning Analytics our target learner's data will be collected from KNOU learning management system and Open University. There are restrictions on the access and use of learner's data due to the privacy policy of universities. It is necessary to use open data from the FUN or to find out methodologies to collect learner's data automatically trying to find a better practical way to overcome this restriction like masking or sampling. The eStat project is a good candidate to collect the data corresponding to deep learning study for statistics education. The eStat can cover students from Elementary, Middle, High School and University level. The QR code related with the eStat is a technical tool collect data for deep learning.

Data cleaning and analysis

The following methods and analysis approaches for Learning Analytics inform faculty, educators, and administrators in higher education who are not experts in LA about the available methods reported in the literature. With the current advent of both blended and online learning opportunities in KNOU, big data and learning analytics are predicted to play a significant role in education in future. When discussing learning analytics methods in education, it is important to provide a background regarding the flow of analytical information. Researchers play a role as they validate and report their research results to inform stakeholders of best practices.

Predication

Predication involves developing a model that uses both a predicted variable and predicator variables. A predicted variable represents a particular component of the data, whereas predicator variables consist of a combination of other data elements. Researchers classify predication into three categories known as classification, regression, and density estimation. Baker (2010) described the three categories as classification methods with the use of decision trees, logistic regression, and support vector machine regression.

Clustering

The use of clustering becomes most valuable when the categories within a group are unknown. How appropriate the set of clusters is may be evaluated by how well the set of clusters fits the data. By dividing a collection of data into logical clusters, researchers can assess how cluster sets explain the meaning of the data.

Relationship mining

The method of relationship mining focuses on the goal of discovering relationships between variables in a set comprised of a large number of variables. Forms of relationship mining may include learning which variables are related to a single variable or discovering what is the strongest relationship between two variables. Two criteria are necessary for relationship mining: statistical significance and interestingness (Baker, 2010).

Discovery with models

In the next method known as discovery with models, the goal is to develop a model using one of the following methods: predication, clustering, or knowledge engineering. Knowledge engineering uses human reasoning for model development. When using the discovery with models method, a prediction model influences a model's generalization across different contexts (Baker, 2010). Separation of data for use in the process of human judgment. Researchers classify the separation of data for use in the process of human judgment method as a visualization method, in which educational data have a particular structure and meaning rooted within that structure. This method possesses two distinct goals identification and classification.

Identifying target courses

An initial benefit that evolves from using big data analysis in education is the ability of educational institutions to identify targeted courses that more closely align with student needs and preferences for their program of study. By examining trends in student enrollment and interests in

various disciplines, institutions can focus educational and teaching resources in programs that maximize student enrollment in the most needed areas of study.

Curriculum improvement

Using big data allows instructors to make changes and adjustments to improve curriculum development in the educational system, such as in the use of curricular mapping of data (Armayor & Leonard, 2010). Through the analysis of big data, educators can determine weaknesses in student learning and comprehension to determine whether or not improvements to the curriculum may prove necessary.

Student learning outcome, behavior, and process

Another key benefit of big data and text mining focuses on the ability of schools and instructors to determine student learning outcomes in the educational process as well as determine how to improve student performance (Bhardwaj & Pal, 2010). Researchers noted that the use of educational data mining contributed to positive results in the learning process (Al-Shammari, Aldhafiri, & Al-Shammari, 2013). Analysis of the data can help educators understand the student learning experience through learner interactions with technology tools such as e-learning and mobile learning (Hung & Zhang, 2012). Use of big data also reveals learning behavior, the impact on adaptive learning, and level of persistence (DiCerbo, 2014) in the learning process.

Personalized learning

Arnold and Pistilli (2012) discussed an early intervention system that demonstrates the benefits and power of learning analytics. As an example, Course Signal provides students with real-time feedback. The components of students' grades, demographic characteristics, academic background, and demonstrated effort are all addressed. The system employs a personalized email and a stoplight, specific color method to indicate progress or lack thereof. Using learning analytics, the concept of personalized learning reveals student success.

Improved instructor performance

The use of data provides an opportunity to improve instructor development so that instructors are better prepared to work with students in a technological learning environment. Through the acquisition of data generated from instructor usage of technology and research tools in online libraries (Xu & Recker, 2012), analysts can determine online behaviors by educators. Therefore, use of this information can help identify areas in need of improvement by the instructor to facilitate enhanced instructor-student interactions in the educational environment.

MODELS AND STATISTICAL ANALYSIS

Eight predictive models including random forest (RF), support vector machines (SVM), shrunken centroid (SC), linear discriminant analysis (LDA), k-nearest neighbor (kNN), logistic regression (logistic), and artificial neural networks (ANN) which is Deep Learning were trained and their performances compared. All models were run in R (version 2.3.0, downloadable from http://cran.cnr.berkeley.edu) except for ANN, which was run in STATISTICA (version 7.1, Statsoft, Inc, Tulsa, OK). Model training was performed on a randomly selected subset of patients and testing on the remaining learners' database. The primary approach was to use the selected explanatory variables to predict the response variable. In addition, for predicting learner's sustainability, the predicted value of "source of sustainability" and selected input variables were utilized to predict the drop out of a learner. Categorical variables were changed accordingly to indicator variables.

In a practical example of analysis we can analyze the behavior of the eStat users in the following way:

Elementary school:	Which kinds of graphs do they use frequently, and why
	Are they able to use graphs for comparison?
Middle school:	Can they understand the measure of central tendency?
	Are they able to compare several groups?
High school:	Can they understand the Binomial and Normal distributions?

University:	Do they know the meaning of the Central Limit Theorem? Confidence Intervals? Testing hypotheses?

Model Evaluation Step

Ten runs of 10-fold cross validation (CV) were performed for each iteration to obtain a reliable result with low mean square error (MSE) and bias. The MSE is a function of bias and variance and when the estimator is unbiased, MSE reduces to variance because the bias term in MSE becomes zero. Each run of cross validation is comprised of independent training and testing database, where 90% of the data is put in the training set and the remaining 10% of the data is put into the test set. For every 10-fold CV, the following statistics were calculated: sensitivity (SN), specificity (SP), accuracy (ACC: the sum of correct predictions divided by total predictions), positive predictive value (PPV: probability that the patient is truly positive given a positive prediction), and negative predictive value (NPV: probability that a patient is truly negative given a negative prediction). For each classification model, statistical results of 10 repetitions of 10-fold CV were averaged and reported. ROC curves were calculated and areas under the curve were compared for each of the eight models. The Mann-Whitney statistic was calculated, which is equivalent to the area under an ROC curve.

Model Validation Step

After models had been trained using a 70% of learner's database, an unseen before 30% learner's database was utilized as a test set to validate the model. Accuracies were compared for both the evaluation step and the validation step for testing data. For the model validation let's check error rate using DLA model for new data and make visualization of significant covariates and critical value for forecasting learner's performance and give an alarm for sustainability.

DISCUSSION

Using Deep Learning Analysis we can find the optimal learning management model for the variety of educational demand and for personalization of instructors and students. We recommend the following:

1. *Support services for instructors* – significantly impacts the quality of ODL. Especially, support services to facilitate course management, interaction between instructors/mentors and learners and interaction among learners becomes very important.
2. *Support services for learners* – is necessary to support announcements from instructors, discussion, help desks for Q&A and activities in online learning communities. In any case, it is the basic principle to provide appropriate feedback and incentive to learners for their learning activities.
3. *Support services for improvement of basic skills of learners* – as online learning utilizes computers and the internet for providing learning activities, if a learner does not have basic skills to operate computers and the internet, it is difficult to ensure the quality of online learning the learner experiences to provide support to improve skills if the skill level of the learner is found unsatisfactory.
4. *Support services for providing appropriate learning environments* – which is important to provide online learning courses appropriate for the learners' existing learning environment.

REFERENCES

Ali, L., Hatala, M., Gašević, D., & Jovanović, J. (2012). A qualitative evaluation of evolution of a learning analytics tool. *Computers & Education*, *58*(1), 470-489.

Al-Shammari, I., Aldhafiri, M., & Al-Shammari, Z. (2013). A meta-analysis of educational data mining on improvements in learning outcomes. *College Student Journal*, *47*(2), 326-333.

Alsheikh, M. A., Niyato, D., Lin, S., Tan, H. P., & Han, Z. (2016). Mobile big data analytics using deep learning and apache spark. *IEEE Network*, *30*(3), 22-29.

Armayor, G. M., & Leonard, S. T. (2010). Graphic strategies for analysing and interpreting curricular mapping data. *American Journal of Pharmaceutical Education*, *74*(5), 81.

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 267-270).

Baker, R. S. J. D. (2010). Data mining for education. *International encyclopedia of education*, *7*(3), 112-118.

Bhardwaj, B. K., & Pal, S. (2010). Data Mining: A prediction for performance improvement using classification. *International Journal of Computer Science and Information Security*, *9*(4), 136-140.

Breiman, L. (2001). *Machine Learning*. Springer.

Business Wire (2018, May 10). AI in Education Market. Article retrieved from https://www.businesswire.com/news/home/20180510005723/en/AI-Education-Market---Forecast-2023-Market

Cho, Y., & Saul, L. K. (2009). Kernel Methods for Deep Learning. In *Advances in neural information processing systems 22*, (pp. 342-350).

DiCerbo, K. (2014). Assessment and teaching of 21st century skills. *Assessment Education: Principles, Policy & Practice*, *21*(4), 502-505.

Hung, J. L., & Zhang, K. (2012). Examining mobile learning trends 2003-2008: A categorical meta-trend analysis using text mining techniques. *Journal of Computing in Higher Education*, *24*(1), 1-17.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1), 1.

Scheffel, M., Drachsler, H. Stoyanov, S., & Specht, M. (2014). Quality indicators for learning analytics. *Journal of Educational Technology & Society*, *17*(4), 117-132.

Vieira, C., Goldstein, M. H., Purzer, Ş., & Magana, A. J. (2016). Using learning analytics to characterize student experimentation strategies in engineering design. (2016). *Journal of Learning Analytics, 3(3)*, 291–317. http://dx.doi.org/10.18608/jla.2016.33.14

Xu, B., & Recker, M. (2012). Teaching Analytics: A clustering and triangulation study of digital library user data. *Educational Technology & Society*, *15*(3), 103-115.