VNU Journal of Science: Education Research, Vol. 37, No. 4 (2021) 19-26

VNU Journal of Science: Education Research

Journal homepage: https://js.vnu.edu.vn/ER

Review Article

The Applications of Machine Learning in Education Science Research

Nguyen Thi Kim Son[1,*], Bui Thi Thanh Huong[2]
Chu Cam Tho[3], Pham Tuan Anh[1], Nguyen Quoc Tri[4]

[1]*Hanoi Metropolitan University, 68 Duong Quang Ham, Quan Hoa, Cau Giay, Hanoi, Vietnam*
[2]*VNU University of Education, 144 Xuan Thuy, Cau Giay, Hanoi, Vietnam*
[3]*The Vietnam Institute of Educational Sciences, 101 Tran Hung Dao, Cua Nam, Hoan Kiem, Hanoi, Vietnam*
[4]*Hanoi National University of Education, 136 Xuan Thuy, Cau Giay, Hanoi, Vietnam*

Received 07 August 2021
Revised 17 November 2021; Accepted 17 November 2021

Abstract: The article presents an overview of the application of machine learning techniques in education science research. The research process shows the use of technology in learning and teaching, collecting information, analyzing and processing data to provide high-accuracy answers or advice in solving educational issues is the trend and strength in education science research. Through this, the authors make recommendations on some research directions in the field of education approaching international publications.

Keywords: Machine learning, data science, education science, international publication.

1. Introduction

Today humanity has entered the era of technology-based creativity. The Industry 4.0 is shaped inseparably with data and data analysis, which poses challenge to organizations, individuals, scientists, researchers, managers and organizers, etc. in handling of data to improve efficiency and capicity of activities and to reduce the risks. For the operation of any agency or unit, data is considered an asset, a core part of strategic activities which brings about great value in organization management and increasing competitiveness. The characteristics of data in the digital age are of very large volume, complex structure, fast change, so techniques have to be further developed to respond to new data analysis needs, transforming data into information and continuing to transform information into operations and strategies for helping leaders and units make important decisions in managing and organizing. In addition, the data will provide sufficient and quantitative scientific arguments for scientists and researchers to make judgments with high accuracy,

* Corresponding author.
E-mail address: ntkson@daihocthudo.edu.vn

https://doi.org/10.25073/2588-1159/vnuer.4562

contributing to properly solving research problems. It can be affirmed that the use of technology in combination with appropriate research methods is the orientation of modern science and technology activities where science and technology have been proven to become a direct productive force.

Over the past two decades there have been significant advances in the field of machine learning. This field has become popular as the method of developing Virtual reality software for computer vision, speech recognition, natural language processing, robot control and other applications accompanied with the trend of using technology in education and training, in which new types of training in data science and artificial intelligence can be considered examples for machine learning research.

With the positive impact of the increase in the amount of educational data through digitization, there are quite a few areas where machine learning can positively affect education. It can be affirmed that this is an inevitable trend which prove for the development of education and training associated with technology in the context of Industry 4.0.

The article shared an overview of the application of machine learning techniques in education science research and some recommendations on some research directions in the field from education approaching international publications were proposed for the context of educational research in Vietnam

2. Literature Review

2.1. Artificial Intelligence and Machine Learning

Machine learning (ML) is the study of computer algorithms that can improve automatically through experience and by the use of data [1] Machine learning is an area of artificial intelligence that deals with the study and construction of techniques that allow systems to "learn" automatically from data to solve specific problems. Machine learning technology can be considered an effective solution for data mining in the current global digital transformation context. Therefore, the research and application of machine learning techniques in education science is very urgent and it should be developed and organized in a methodical way to meet the requirements of digital transformation in schools, in education and training in general today.

In the field of Machine Learning an advantage to this technology is that computers do not need to be programmed explicitly and specifically, computers are fully capable of changing and improving factors about algorithms, or in other words computers are approaching artificial intelligence (AI) technology.

Artificial intelligence is an important technology of the world's Industry 4.0. This technology is widely applied in many fields such as: economy, culture, society, education,... AI is applied not only for modern scientific and technological facilities but also gradually for all areas of life.

2.2. Machine Learning Research in Education

One of the first uses of machine learning in education was helping quizzes and tests go from multiple choice to fill-in-the-blank. The evaluation of students' free-form responses is based on Natural Language Processing (NLP) and machine learning. Various studies on the effectiveness of automated scoring have shown better results than human scoring in some cases. Furthermore, automatic scoring provides more scores faster than humans, which makes it useful for creative assessment.

A few years ago, prediction was considered an application of machine learning in education. A study conducted by Kotsiantis (2012) [2] presented a new case study describing the emerging field of machine learning in education. In this study, student-specific data and grade data were mined as a dataset for a regression machine learning method used to predict students' future academic performance. Likewise, a number of projects have been conducted including one that aims to develop a predictive model that can be used by educators, schools and policy makers to predict the risk of

students dropping out of school. IBM's ChalapathyNeti shared IBM's vision for Smart Classrooms using cloud-based learning systems that can help teachers identify students at highest risk of dropping out and observe why they havedifficulty, as well as provide details on the interventions needed to overcome their learning challenges.

Supervised learning is based on learning from a set of labeled examples in the training set so that unlabeled examples in the test set can be identified with the highest possible accuracy (G. Erik, 2014) [3]. This model of learning is very efficient and it always finds solutions to some linear and non-linear problems such as classification, vegetation control, forecasting, prediction, robotics and many other matters (Sathya and Abraham 2013).

Some existing works have focused on supervised learning algorithms such as Naive Bayesian Algorithm, Association Rule Mining, Artificial Neural Network (ANN)-based algorithms, Logistic Regression, CART, C4. 5, J48, (BayesNet), SimpleLogistics, JRip, RandomForest, Logistic Regression Analysis, ICRM2 for classification of school dropouts (Kumar et al., 2017). However, according to classification techniques, Neural Networks and Decision Trees are two methods that are widely used by researchers to predict student performance (Shahiri et al., 2015). The advantage of neural network is that it is capable of detecting all possible interactions between predictor variables (Gray et al., 2014) and can also perform complete detection even inthe complex nonlinear system between dependent and independent variables (Arsad, PauziahMohdBuniyamin, Norlida Manan, 2013), while decision map has been used because of its simplicity and ease of exploration for discovery of smaller or larger data structures and value prediction (Natek and Zwilling, 2014) [4].

Unlike supervised learning algorithms, unsupervised learning algorithms are used to identify hidden patterns in unlabeled input data. It refers to the ability to learn and organize information without signaling errors and to be able to evaluate potential solutions. Sometimes the lack of direction for the learning algorithm in unsupervised learning can be beneficial, as it allows the algorithm to find patterns that have not been considered before (Sathya and Abraham, 2013) [5].

Matrix analysis is a clustering machine learning method that can fit several variations (Yang et al., 2014) [6]. The study presented by Hu and Rangwala (2017) describes matrix analysis. In Elbadrawy et al., (2016) [7], two classes of methods for building predictive models have been presented. The purpose of the research is to facilitate degree planning and identify who is at risk of failing or dropping a class. The first layer builds the model using linear regression and the second layer uses matrix analysis. Regression-based methods describe course-specific regression and personalized multilinear regression while methods based on matrix analysis incorporate a standard matrix decomposition approach. The mentioned approach is applied on dataset generated from George Mason University (GMU) transcript data, University of Minnesota (UMN) transcript data, UMN LMS data and MOOC data of Stanford University. One limitation of the matrix decomposition method is that it ignores the sequence in which students took different courses. In addition, the latent representation of a course can be influenced by a student's performance in subsequent courses.

Furthermore, the study presented in the work of Iam-On and Boongoen (2017) [8] has proposed a new data transformation model, which is built on the summary data matrix of combinatorial clusters based on link. The aim of the research was to establish the clustering method as a practical guide to explore the types and characteristics of students. This was done by using an education dataset obtained from an operational database system at Mae Fah Luang University, Chiang Rai, Thailand. Like some existing dimensionality reduction techniques such as Principal Component Analysis and Core Principal Component Analysis, this method aims to achieve high classification accuracy by transforming the original data into a new form. However, the common limitation

of these new techniques is that it requires time complexity, so it may not scale well to a very large data set. Although the worst-case review time is not strictly for a time-intensive application, it can be an attractive candidate for quality research, such as identifying high school students at risk of failing.

Deep Neural Network (DNN) is an approach based on Artificial Neural Network with many hidden layers between input and output layers (Deng and Yu, 2014) [9] while the Probability Graph Model (PGM) combines probability theory and graph theory to provide a compact graph-based representation of general probability distributions exploiting conditional independence among random variables (Pernkopf et al., 2013). Similar to shallow ANN, DNN can modelize complex non-linear relationships (Ramachandra and Way, 2018) [10]. Various deep learning architectures such as Recurrent Neural Networks (RNNs) and other probabilistic graphical models such as Hidden Markov Models (HMMs) have been used for the dropout problem (Fei and Yeung 2015) [11].

The study presented by Fei and Yeung (2015) was considered two temporal models - the state space model and the cyclic neural network. These approaches have been applied in two MOOC datasets, one provided on the Coursera platform, called "Culinary Science" and the other on the edX platform, called "Gender" Introduction to Java Programming". The purpose of the research was to identify students at risk of dropping out. State space model describes two variants of Input Output Hidden Markov Model (IOHMM) with continuous state space while recurrent neural network describes RNN and RNN cells with short term memory cells lengths (LSTMs) are hidden units. IOHMM is recommended for learning problems involving sequential structured data. Since it is derived from the HMM, it has learned to map the input string to the output string. Furthermore, unlike the standard discrete-state HMM, the state space in the IOHMM formula is described as continuous, so the state space can carry more representation than enumeration of states. Moreover, unlike feed-forward neural networks such as multilayer Perceptrons, recurrent neural networks allow network connections to form cycles.

Many schools have now begun to create personalized learning experiences through the use of technology in the classroom. Thanks to the advancement of the amount of data collected, machine learning techniques have been applied to improve the quality of education including areas related to learning and content analysis (Lan et al., 2014), knowledge seeking (Yudelson et al., 2013) reinforcement of learning materials (Rakesh et al., 2014) [12] and early warning systems (Beck and Davidson 2016) [13]. The use of these techniques for educational purposes is a promising area for the development of methods to explore data from educational institutions that compute and discover meaningful patterns (Nunn et al., 2016) [14].

In Vietnam, some scientists have initially conducted research on the application of machine learning in education science. Having taken the advantage of the superior features of the Mymedialite system, (Listen, 2013) [15] they built a method to predict student performance. The authors have shown that it is only necessary to build a program that reads the data, checks and converts them to a format suitable for the prediction algorithm, configures the input parameters of the algorithm, calls the built-in functions in the library to train and predict the results, and finally save the prediction data to the database so that it can be exploited to the needs of the application system. However, the authors also recommend that there are some problems when switching to the system of suggesting subject selection, which is to pay attention to logic, pedagogy and specialized orientationbecause the problem of predicting learning outcomes is solved by an approach similar to the ranking problems in RS. This is the matter that needs further research.

Another study using two data mining algorithms Naïve Bayes and Logistic Regression also gave some positive results in predicting learning outcomes and predicting forced withdrawal (Uyen and Tam, 2019) [16]. With this algorithm, it is possible to accurately indicate which students need to study hard to reduce the risk of being forced to stop studying.

With 18 data fields but focusing mainly on 2 main attribute fields namedgender and cumulative score, the authors (Sang, Dien, Nghe and Hai, 2020) [17] have proposed a method to predict students' learning results by deep learning techniques to exploit the database in the student management system at Can Tho University. After collecting data, the authors conduct analysis, select suitable attributes, preprocess the data, design and train the MLP network. With the design and training of multi-layer neural networks, the results obtained on predicting student learning outcomes are the initial results in applying machine learning and deep learning techniques to support management process of training activities in the university.

3. Methodology

By using some methods for literature review such as: narrative review, descriptive review, scoping reviews, systematic reviews, searching the extant literature, some findings about literature review of applications of machine learning in education research were proposed.

Literature review was implimented through steps: screening for inclusion, assessing the quality of primary studies, analyzing and synthesizing data.

4. Findings

Machine learning is a technology that is applied in many fields. Although Machine Learning technology also has a similar feature to technologies with limitation to implementation of solving research problems, it is necessary to emphasize that the fields in Machine Learning are very broad, especially in functions of education science.

Nowadays the education context has changed greatly when the learning conditions of learners are improved with investment both at the national, school and learner levels. Technology has become a productive part of the educational process. In addition, individual learning needs are focused. Therefore, pedagogical research is being redirected to in-depth study of learner behavior to establish individual learning programs; at the same time it exploits big data of learners to diagnose and reorient the learning process of learners in particular, and manage/operate the educational process in general. It is also the content that machine learning can be applied to in education science research. In this article, we focus on in-depth analysis of some problems applying machine learning in supporting education scientific research.

4.1. The Problem of Image Processing in Education

The education trend of the 21st century is open education and mass education, so in addition to the theoretical knowledge imparted academically, it is necessary to have the participation of images, which contributes to the development of the society to support for the theoretical knowledge that needs to be conveyed and this is also relevant to the issue of open education and mass education in the current context.

4.2. The Problem of Text Analysis and Data Mining in Education

With the trend of open education and mass education this is the basis for the formation and construction of open learning resources, which requires convenient technology algorithms to solve these two problems. The problem with open learning materials is to discover and gather databases of scientific and technological knowledge; besides there is a need for an AI-based technology, artificial intelligence to support learners as well as teachers in analyzing texts based on the needs and goals of teachers and learners.

Text analysis is the process of extracting or classifying information from text.

The problem of information extraction (Information Extraction): this is the problem of manipulating the algorithm most used by teachers and learners. From a specific database, this operation can help teachers and learner extract the information and search fields that are suitable for the field and the knowledge that teachers and learners need to be provided.

Data mining is the process of discovering valuable information or making predictions from data. This is an overarching problem, but now databases are in the form of Big data with very large data sources, which will be very difficult for learners as well as teachers in education if they can't do data mining proficiently, then it will be also difficult to find the right data.

5. Discussion

About *the Problem of Image Processing in Education*, Machine learning can process images based on the following methods and applications such as analyzing information from images and from an image needing processing, Machine Learning technology can handle a number of following problems:

Firstly, the operation of tagging images (Image Tagging). This is a very familiar operation that appears on popular social networks such as Facebook, Instagram, Tiktok. This is a function based on the broadcasting algorithm and self-recognize the individual's faces in line with the individual's images and faces stored in databases. This algorithm is manipulated based on the provided database and then automatically finds similar photos of an individual that have been used before. Using this image tagging function is very possible in the field of education based on additional jobs for teachers such as attendance, class list management, as well as being convenient for learners when being accepted and taking the information of the class.

Secondly, the operation of character recognition (Optical Character Recognition), which is also a very familiar operation. Now there are many applications on smartphones that help users to store the following documents in jpg or pdf file format. The character recognition algorithm from machine learning is an upgrade of technology, this is a very important field in education, especially in the context of education that is promoting digitization, digital transformation, having a Machine Learning algorithm with the ability to recognize characters, documents in the form of characters are presented, recognized by this algorithm and converted to digitized form for use and storage is considered very useful. This is an algorithm that greatly contributes to both learners and teachers in the problem of storing information and communicating knowledge between teachers and learners in both face-to-face and online interactions.

About *the Problem of Text Analysis and Data Mining in Education*, the problem of manipulating the algorithm most used by teachers and learners, was considered through the below issues:

Firstly, the algorithm detects anomaly (Anomaly detection); this is the method by which the algorithm detects anomalies, such as cheating in the learning process, or at a higher level, it detects anomalies in the research and development (R&D) of a science and technology activity in a university. To be able to detect anomalies, it is necessary to mine data with anomalous properties and compare it with standard values so as to synthesize and make an assessment of the operation. This is a very necessary and essential algorithm for teachers and learners.

Secondly, the algorithm detects the rules (Association rules): the data mining of teachers and learners often takes place many times, from which the algorithm will build a database of trends in science and technology needing to search, then it will synthesize search rules as well as frequently searched fields for teachers or learners, and finally AI technology will make

predictions about search trends as well as propose scientific fields and necessary knowledge in accordance with the search trends of teachers and learners.

Thirdly, grouping algorithm (grouping) is also an important algorithm. Grouping operation is the operation often used by teachers in dividing students in the class into groups based on common characteristics as well as the appropriate field of study. With the background of AI technology and database of learners, grouping will be easier for the teacher to manipulate and it is also suitable to the characteristics of the learners.

Fourthly, prediction - this is an algorithm with predictive nature. It can be confirmed that predictive research is a difficult type of research in science and technology. In teaching activities, teachers need to do experiments to verify the responses of those parameters in practical conditions. The use of AI and this algorithm contributes to predicting research results, ensuring cost and safety for teachers and learners.

6. Conclusion

For the most part the application of machine learning in particular and data mining in general in education research is various. However, domestic research in this area is still quite limited. One of the main reasons is that the digital transformation in education in Vietnam is relatively slow compared to other countries in the world. The collection of digital data, digital transformation of contents in education in general and in schools are being carried out in initial steps. In addition, data mining algorithms and machine learning techniques are increasingly developed, the choice of which algorithm is suitable for logic, the requirements of educational problems is an issue that should be further promoted in research. This is the initial approach for the birth and growth of a new research trend - the application of artificial intelligence (AI) in education.

Reference

[1] Mitchell, Tom, Machine Learning, ISBN 0-07-042807-7, OCLC 36417892, New York: McGraw Hill, 1997.

[2] S. B. Kotsiantis, Use of Machine Learning Techniques for Educational Proposes: A Decision Support System for Forecasting Students' Grades, Artificial Intelligence Review, Vol. 37, No. 4, 2012, pp. 331-344, https://doi.org/10.1007/s10462-011-9234-x.

[3] G. Erik, Introduction to Supervised Learning, Data Mining and Knowledge Discovery Handbook, 2014, pp. 149-164, https://doi.org/10.1007/0-387-25465-X8.

[4] P. M. Arsad, N. Buniyamin, J. A. Manan, A Neural Network Students' Performance Prediction Model (NNSPPM), IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA), 2013, pp. 1-5.

[5] G. Gray, C. McGuinness, P. Owende, An Application of Classification Models to Predict Learner Progression in Tertiary Education, IEEE International Advance Computing Conference (IACC), 2014, pp. 549-554, https://doi.org/10.1109/IAdCC.2014.6779384.

[6] M. Kumar, A. J. Singh, D. Handa, Literature Survey on Educational Dropout Prediction, J. Education and Management Engineering, Vol. 2, 2017, pp. 8-19, https://doi.org/10.5815/ijeme.2017.02.0.

[7] S. Natek, M. Zwilling, Expert Systems with Applications Student Data Mining Solution Knowledge Management System Related to Higher Education Institutions, Expert Systems with Applications, Vol. 41, 2014, pp. 6400-6407, https://doi.org/10.1016/j.eswa.2014.04.02.

[8] A. M. Shahiri, W. Husain, N. A. Rashid, A Review on Predicting Student's Performance Using Data Mining Techniques, Procedia Computer Science, Vol. 72, 2015, pp. 414-422, https://doi.org/10.1016/j.procs.2015.12.157.

[9] R. Sathya, A. Abraham, Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification, International Journal of Advanced Research in Artificial Intelligence, Vol. 2, No. 2, 2013, pp. 34-38, https://doi.org/10.14569/IJARAI.2013.020206.

[10] D. Yang, M. Piergallini, I. Howley, C. Rose, Forum Thread Recommendation for Massive Open Online Courses, Proceedings of the 7th International Conference on Educational Data Mining (EDM), 2014, pp. 257-260.

[11] A. Elbadrawy, A. Polyzou, Z. Ren, M. Sweeney, G. Karypis, H. Rangwala, Okay Predicting Student Performance Using Personalized Analytics, Computer, Vol. 49, No. 4, pp. 61-69, https://doi.org/10.1109/MC.2016.119. 12.

[12] M. Fei, D. Y. Yeung, Temporal Models for Predicting Student Dropout in Massive Open Online Courses, IEEE International Conference on Data Mining Workshop (ICDMW), Vol. 2, No. 15, 2015, pp. 256-263, https://doi.org/10.1109/ICDMW.2015.174.

[13] I. O. Natthakan, T. Boongoen, Generating Descriptive Model for Student Dropout: A Review of Clustering Approach, Human-centric Computing and Information Sciences, Vol. 7, No. 1, 2017, pp. 1-24, https://doi.org/10.1186/ s13673-016-0083-0.

[14] L. Deng, D. Yu, Deep Learning: Methods and Applications, Foundations and Trends® in Signal Processing, Vol. 7, No. 3-4, 2014, pp. 197-387.

[15] V. Ramachandra, K. Way, Deep Learning for Causal Inference, 2018.

[16] M. Fei, D. Y. Yeung, Temporal Models for Predicting Student Dropout in Massive Open Online Courses, 2015 IEEE International Conference on Data Mining Workshop (ICDMW), 2015, pp. 256-263, https://doi.org/10.1109/ICDMW.2015.174.

[17] A. Rakesh, M. Christoforaki, S. Gollapudi, A. Kannan, K. Kenthapad, A. Swaminathan, Mining Videos from the Web for Electronic Textbooks, Microsoff Research, Yudelson, MV, Koedinger, KR and Gordon, GJ, Individualized Bayesian Knowledge Tracing Models, 2014.

[18] H. P. Beck, W. D. Davidson, Establishing an Early Warning System: Predicting Low Grades in College Students from Survey of Academic Orientations Research in Higher Education, 2016.

[19] S. Nunn, J. T. Avella, T. Kanai, M. Kebritchi, Learning Analytics Methods, Benefits, and Challenges in Higher Education: A Systematic Literature Review, Online Learning, Vol. 20, No. 2, 2016, pp. 13-29, https://doi.org/10.24059/olj.v20i2.790.

[20] N. T. Nghe, T. Q. Dinh, University Admissions Support System, Science Magazine Can Tho University, 2015, pp 152-159, https://sj.ctu.edu.vn/ql/docgia/tacgia711/baibao31511.html (inVietanmese).

[21] N. T. Uyen, N. M. Tam, Predicting Student Learning Outcomes by Data Mining Techniques, Scientific Journal - Vinh University, No. 48, 2019, pp. 68-73 (in Vietnamese).

[22] L. H. Sang, T. T. Dien, N. T. Nghe, N. T. Hai, Predicting Learning Outcomes by Deep Learning Techniques with Multilayer Neural Networks, Can Tho University Journal of Science, Vol. 56, No. 3, 2020, pp. 20-28, https://doi.org/10.22144/ctu.jvn.2020.049 (in Vietnames).