Martin BORDES

July 2023

Abstract

In the recent years, artificial intelligence is gaining more and more in importance in our society and will probably impact how we do things in a lot of area, if not all. Education can be seen as the domain at the center of our society and should be no exception. The goal of this theoretical work is to study how artificial intelligence is impacting and will change education on several different topics.

We will first look at the main topic when it comes to education and machine learning : the prediction of student performance, including student retention. Then, we will talk about other applications of AI in education such as automatic question generation and assessment.

Contents

1 Introduction

In the last decades, education has known a lot of changes, following those of the society. These changes are often described as four distinct phases. During the first phase, that lasted for a very long time, teaching was authoritarian. It was a a teacher-centered system where the student was just a recipient. Any technology was forbidden. The second phase is a transition phase. Communication and collaboration are emerging and some start to talk about learning instead of teaching. This is also an exam-based approach that puts memorization of knowledge at the center. Then, the third phase coincides with the rapid development of internet. The student is researching and the teacher is now a facilitator. There are more flipped classrooms then more dialogues. The system is now student-centered. We are currently in this Education 3.0, especially in higher education. The future, the fourth phase, will probably leave plenty of space to innovation. That will also probably be the end of classical classrooms and lectures. Learning will become an online activity. This will help to create a lot more data that will allow us to innovate more easily, especially in machine learning.

Artificial intelligence will probably be one of the bases, if not the main base, of the future of education. Solutions will be implemented at each phase of learning from designing lectures to assessing exams. All the chain will be specifically adapted to each student to optimize his chances of succeeding. Researchers are already studying such solutions, providing interesting results in several domains of education.

2 The main topic : Student retention and performance

The first idea that comes in mind when thinking about education and machine learning is probably the prediction of a student's level or particularly his future grades. It could help identify weaknesses specific to each student in order to help them efficiently. It is also possible to group students depending on their performance to attend courses adapted to their level. However, this approach can be controversial because it could prevent to give all students the same chances to succeed in the future. In fact, these types of problems contains one that is very crucial : improving student retention.

This task might be the most important when dealing with education and machine learning. This can have several benefits both for students and for schools. The goal is to help struggling students to succeed by providing them a particular attention. For example, they could benefit from extra hours in classroom, guided by mentors. This could really help a lot of students in going through a difficult moment in their academic career. Moreover, it could also profit to schools and universities. Indeed, dropouts are an important cost for these facilities. A lot of money is spent on each student, especially in some countries, and a student dropping out is then a waste of resources. Furthermore, school rankings are becoming increasingly important these days and the drop-out rate is one of the main factors potential future students might want to consider. It could also influence possible investors. The goal is to identify "at risk" students as early as possible. However, a model will be obviously more precise later in the semester when it can use more information about homework or intermediate tests. Here, we will not distinguish failures from dropouts since they have the same consequences on students and schools and consists in a similar problem from a machine learning point of view.

2.1 Classical methods

Djambic et al. (2016) implemented such a program in the University College Algebra in Zagreb, Croatia. The focus was put on introductory programming course, which has a significant failure rate, particularly for students whose major is not computer science. They used as inputs the first colloquium of the semester, the first quiz, the first homework, whether is this a second time student has enrolled in this course and whether the student has attended the first colloquium. They compared three models of logistic regression : a linear model, a quadratic model with only quadratic members and a quadratic model with all members :

$$h_{1,\theta}(x) = S(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \theta_5 x_5)$$

$$h_{2,\theta}(x) = S(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \theta_5 x_5 + \theta_6 x_1^2 + \theta_7 x_2^2 + \theta_8 x_3^2 + \theta_9 x_4^2 + \theta_1 0 x_5^2)$$

$$h_{3,\theta}(x) = S\,(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \theta_5 x_5 + \theta_6 x_1^2 + \theta_7 x_2^2 + \theta_8 x_3^2 + \theta_9 x_4^2 + \theta_1 0 x_5^2 + \theta_{11} x_1 x_2 + \theta_{12} x_1 x_3 + \theta_{13} x_1 x_4 + \theta_{14} x_1 x_5 + \theta_{15} x_2 x_3 + \theta_{16} x_2 x_4 + \theta_{17} x_2 x_5 + \theta_{18} x_3 x_4 + \theta_{19} x_3 x_5 + \theta_{20} x_4 x_5\,)$$

The best performance was achieved by the quadratic model with only quadratic members. They considered students as "at risk" when the probability of failure was over 40%. Those students were sent to additional classes. The misclassification error was around 19%, which is not very good. However, the precision was 67% and the recall 92%, which means that they were sending a lot of successful students to additional classes and missed little "at risk" students. In fact, the recall is probably more important than the precision here because having good students attending extra hours is not really a problem.

This model did not deal with one of the main problem concerning this task : the potential imbalanced data. Indeed, a large majority of student is usually passing a course. This creates a dataset imbalanced with respect to the target and can affect the performance of the model by having a very poor accuracy on the minority class. Thammasiri et al. (2014) looked into this problem. They compared three data-balancing techniques using four types of classifiers. The dataset they used was large and rich. It contained 21,654 examples on 34 variables of students enrolled as freshmen between 2005 and 2011. The features were related to student's academic, financial and demographics characteristics. The four classification methods were artificial neural networks, support vector machines, decision trees and a logistic regression. The first data-balancing technique is the random under-sampling (RUS). It consists in randomly remove some majority class examples until the number of samples is nearly equal between both classes. The second technique is the random over-sampling (ROS). It consists in randomly select examples from the minority class and count them as new samples in the dataset until the data becomes roughly balanced. The last method is the synthetic minority over-sampling technique (SMOTE). The process starts by finding the k nearest neighbours from the same class of each sample from the minority class. Then, we randomly select some of them. Finally, we create a new synthetic example along the line between each minority example and each of its selected neighbours. This is repeated until both classes contain nearly the same number of samples. The support vector machines using the SMOTE is the one that has performed the best in term of accuracy (90%) and specificity (96%).

More recently, Palacios et al. (2021) conducted a similar study in Chile. They predicted student retention during their first, second and third year of study using several machine learning algorithms : decision trees, k-nearest neighbors, logistic regression, naive Bayes, random forest, and support vector machines. Their accuracy exceeded 80% in most of the cases but the best result was achieved by the random forest with a score of 88,4% and a false positive rate of 15,1%. As in the last discussed paper, the use of balancing techniques was truly beneficial. For example, applying the SMOTE to the random forest increased the accuracy from 81,8% to 88,4% and more importantly decreased the false positive rate from 87,9% to 15,1%.

Most of the time, universities and school make use of academic information to predict dropouts, as we seen with Djambic et al. (2016). However, Deepti et al. (2021) showed adding non-academic variables, mainly demographic, can improve the model performance. They took into account the year of birth, the gender, the month of birth, the age at the time of admission, the Indian state (the study takes place in India) as well as the parent annual income. They compared the results of eight models using only the academic parameters and using all parameters. The F1-score increased significantly for all models after adding the demographic features. They all ranged between 78,1% and 79,6% without and between 90,3% and 93,8% with.

2.2 Different approaches

All papers we have seen so far made use of common techniques of classification such as random forests, support vector machines or logistic regressions. Bydžovská (2015) tried a different approach, a collaborative filtering method. This is usually implemented for recommender systems but they applied it in the context of student performance. They used as inputs the grades of past years. Then, they transformed those grades into categorical values. From that, they constructed a similarity matrix for each student with a set of students enrolled in the course in the last two years. It allows to define a score of similarity between with these past students and thus a neighborhood. Finally, the prediction was made using the grades of students in the neighborhood. They used this process for three different tasks : predicting the exact grade, if the student will have a good grade (A,B) a bad grade (C,D,E) or will fail (F,-) and if the student will pass or fail. In these three tasks the performance was similar to that of the pure classification algorithms they studied (Bydžovská and Popelínský (2014)).

Most of the methods we have seen so far used basic machine learning techniques because they were not recent enough. Indeed, in the last years deep learning has gained a lot of interest in almost every application of machine learning and education is no exception. Uddin et al. (2021) compared different methods and the best result was achieved by a convolutional neural network. This is a type of network using a convolutional layer which performs a convolutions on adjacent points. They used 21 academic-related variables as inputs. Moreover, the authors did not predict only the probability of failure or dropout but the exact mark. The network architecture was as follows :

<table>
<tr><th>Operation</th><th></th><th>Data dimensions</th><th>Weights (n)</th><th>Weights (%)</th></tr>
<tr><td>Input</td><td>####</td><td>21 1</td><td></td><td></td></tr>
<tr><td>conv1D</td><td>\|/</td><td>---------------------</td><td>256</td><td>17.4%</td></tr>
<tr><td>relu</td><td>####</td><td>19 64</td><td></td><td></td></tr>
<tr><td>flatten</td><td>|||||</td><td>---------------------</td><td>0</td><td>0.0%</td></tr>
<tr><td></td><td>####</td><td>1216</td><td></td><td></td></tr>
<tr><td>Dense</td><td>XXXX</td><td>---------------------</td><td>1217</td><td>82.6%</td></tr>
<tr><td>Sigmoid</td><td>####</td><td>1</td><td></td><td></td></tr>
</table>

Figure 1: Architecture of the convolutional neural network

3 Other interesting tasks

If predicting students' performance is probably the main task about education and artificial intelligence, it is not the only one. The scope of this area is very large and it can use very different aspects of artificial intelligence. We will now see two other examples of the impact of the rise of machine learning on education : automatic question generation and automatic assessment.

3.1 Automatic question generation

In the way education is currently working, designing tests is a task fully to be performed by the teachers. This is not optimal because it takes a very large amount of time to them to do that, a time that could be seen as lost. Moreover, it is impossible to adapt tests to students this way. The development of automatic test generation systems could be a response to these issues.

The most common aspect of test generation is natural question generation (NQG). The goal is to use a passage of text, an image or other inputs as well as an answer to create a question related to this extract. Chen et al. (2020) proposed such a solution using a reinforcement learning based graph-to-sequence model. Their overall architecture is as follows :

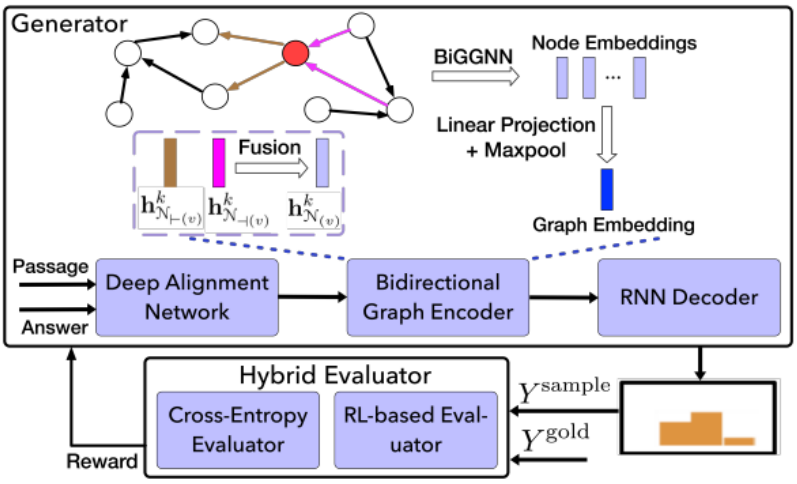

Figure 2: Overall architecture of the model proposed by Chen et al. (2020)

First, the deep alignment network uses passage embeddings and answer embeddings to create a final passage embedding thanks to an attention-based mechanism. Attention allows the network to enhance some parts of the inputs and reduce the impact of others, as a human would do with cognitive attention. A simple example would be to focus specifically on the pixels where a dog is when looking at a dog image.

Then, the final passage embedding is used to construct a graph representing the passage. This graph is passed through a graph neural network that outputs node embeddings. Those objects were build using information about the final passage embeddings and the graph structure. This structure was based either on semantics or on syntax depending on the graph construction technique applied. The node embeddings are finally converted into a graph embedding thanks to linear projection and max pooling. This whole part is the encoder.

The decoder part will then aim to produce the question from the information extracted by the encoder. Here, the authors used a recursive neural network, a type of structure useful for sequence data. More precisely, the model used is a long-short-term-memory decoder that takes the graph embedding as inputs for building the initial hidden states and uses the node embeddings as an attention memory. Finally, for evaluating the question produced by the decoder, the authors use a hybrid method. They combined a classic cross-entropy evaluation and a part based on reinforcement learning.

When this paper was published, it outperformed significantly the existing methods on this task, achieved the state-of-the-art results. However, during the last three years, the development of natural language processing was impressively rapid. Large language models like GPT-4 (OpenAI (2023)) will probably be the base of fine-tuned models for most of NLP tasks and automatic question generation should not be an exception.

3.2 Automatic assessment

After having seen the problem of automatic test generation, the next logic topic is automatic assessment. In fact, the first one is probably more complex because it needs to use fully-generative models or at least encoder-decoder ones, as in the paper we discussed. With automatic assessment, the task can either be a classification problem if we want to determine if an answer is right or more generally a regression problem to assign a score to the student's answer.

Grivokostopoulou et al. (2017) developed a full educational system based on artificial intelligence. As part of it, they created an automatic assessment method. Here, the answers are strings of different sizes. Their mechanism first calculates the differences and similarities between the correct answer and the student's one. Then, it identifies the correctness (to what extent the student gave wrong answers) and accuracy (to what extent the student gave wrong answers). It also takes into account the carelessness errors. Based on all of that, their mechanism finally give a score to the answer. They compared their algorithm to human corrections and both turned out to be very close. The R^2 metric of a linear regression between the pairs of scores was equal to 0.934.

However, there are more difficult tasks than assessing simple string answers. It is even more interesting to learn to grade things such as full essays. Tashu et al. (2022) proposed a solution to this task. Their method is composed of both recurrent neural networks and and convolutional neural networks. The CNN part uses word embeddings to create context features as an essay level thanks to max pooling. Then, the RNN part takes these representations as well as those at the word level to finally output a grade to the essay. The authors compared their solution to state-of-the-art models on eight datasets. They provided the better results for seven of them.

The more advanced task related to automatic evaluation is probably the oral-based assessment. Jiao et al. (2021) developed a system to help Chinese students to improve their English and especially their pronunciation. Their model takes as input a recording of the student speaking in English and can detect the pronunciation errors. A deep belief network is used to extract features from the pre-processed recording. Then, a support vector machine takes those features to classify the input into 6 types of pronunciation errors. The authors reached an accuracy above 80% in detecting raising, lowing and shorting errors. Chen (2022) studied a close task. Indeed, they proposed a model to automatically assess spoken English by Chinese students based not only on their pronunciation but on fluency, emotional expression and rhythm. Here, the approach was multimodal, using both speech and text.

4 Conclusion

As a very central domain, education is one of the most affected by society changes. Artificial intelligence will probably one of those major evolutions in a close future and researchers are already working on AI-related solutions for education. The goal of this work was then to look at those techniques.

The hottest area in this domain is the student performance prediction. In particular, a lot of researchers are working on approaches that aim to predict dropouts and failures. This can benefit both to struggling students by providing them additional help and to schools for which retention is a financial and reputational challenge. We have seen classical machine learning methods as well as more complex or unexpected approaches.

Then, we have talked about other ways for AI to impact education. We studied especially a model able to automatically generate questions based on text passages and answers. Other methods are also able to automatically assess students work, such as written essays or spoken English.

References

Bydžovská, H. and L. Popelínský (2014): "The Influence of Social Data on Student Success Prediction," New York, NY, USA: Association for Computing Machinery, IDEAS '14, 374–375.

Bydžovská, H. (2015): "Student Performance Prediction Using Collaborative Filtering Methods," in *Artificial Intelligence in Education*, ed. by C. Conati, N. Heffernan, A. Mitrovic, and M. F. Verdejo, Cham: Springer International Publishing, 550–553.

Chen, X. (2022): "Automatic Assessment Method of Oral English Based on Multimodality," *Scientific Programming*, 2022, 1–10.

Chen, Y., L. Wu, and M. J. Zaki (2020): "Reinforcement Learning Based Graph-to-Sequence Model for Natural Question Generation," .

Deepti, A., S. Mittal, and V. Bali (2021): "Significance of Non-Academic Parameters for Predicting Student Performance Using Ensemble Learning Techniques," *International Journal of System Dynamics Applications (IJSDA)*, 10.

Djambic, G., M. Krajcar, and D. Bele (2016): "Machine learning model for early detection of higher education students that need additional attention in introductory programming courses," *International Journal of Digital Technology and Economy*, 1, 1–11.

Grivokostopoulou, F., I. Perikos, and I. Hatzilygeroudis (2017): "An educational system for learning search algorithms and automatically assessing student performance," *International Journal of Artificial Intelligence in Education*, 27, 207–240.

Jiao, F., J. Song, X. Zhao, P. Zhao, and R. Wang (2021): "A spoken English teaching system based on speech recognition and machine learning," *International Journal of Emerging Technologies in Learning (iJET)*, 16, 68–82.

Kučak, D., V. Juricic, and G. Đambić (2018): *Machine Learning in Education - a Survey of Current Research Trends*, 0406–0410.

OpenAI (2023): "GPT-4 Technical Report," .

Palacios, C. A., J. A. Reyes-Suárez, L. A. Bearzotti, V. Leiva, and C. Marchant (2021): "Knowledge Discovery for Higher Education Student Retention Based on Data Mining: Machine Learning Algorithms and Case Study in Chile," *Entropy*, 23.

Rastrollo-Guerrero, J. L., J. A. Gómez-Pulido, and A. Durán-Domínguez (2020):

"Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review," *Applied Sciences*, 10.

Tashu, T. M., C. K. Maurya, and T. Horvath (2022): "Deep Learning Architecture for Automatic Essay Scoring," .

Thammasiri, D., D. Delen, P. Meesad, and N. Kasap (2014): "A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition," *Expert Systems with Applications*, 41, 321–330.

Uddin, M. I., S. Li, and T. Liu (2021): "Performance Prediction for Higher Education Students Using Deep Learning," *Complexity*.