The ATENA project: an international contribution to AI in Poland

Agata Giełczyk[1][0000−0002−5630−7461], Alfredo Petruolo[2], Michelangelo Agorini[2], Martyna Tarczewska[1], Karol Hartwig[1], and Gracjan Kątek[1]

[1] Bydgoszcz University of Science and Technology
Bydgoszcz, Poland
agata.gielczyk@pbs.edu.pl
[2] Università degli Studi di Napoli "Parthenope"
Napoli, Italy

Abstract. In this article we present the project ATENA: Artificial inTElligence traiNing progrAmme, which is an ongoing programme for students realized on the Bydgoszcz University of Science and Technology. During the programme foreign students have the opportunity to learn and get involved in the artificial intelligence. They follow some projects concerning the real-life issues and provide some promising results. The aim of this paper is to describe the ATENA project and its contribution to the AI development in Poland.

Keywords: ATENA · artificial intelligence · machine learning

1 Introduction

The Artificial Intelligence has become increasingly popular recently. The European Commission highlighted the significance of the AI development in Coordinated Plan on Artificial Intelligence which is also available online: *Artificial Intelligence (AI) can help us address some of the world's biggest challenges. It can enable doctors to improve diagnoses and develop therapies for diseases for which none exist yet; it can reduce energy consumption by optimising resources; it can contribute to a cleaner environment by lessening the need for pesticides; it can help improve weather prediction and anticipate disasters; and so on. The list is virtually endless. AI will be the main driver of economic and productivity growth and will contribute to the sustainability and viability of the industrial base in Europe. Like the steam engine or electricity in the past, AI is transforming the world*[1]. As the AI is one of the fastest developing element of computer science, there is still not enough AI specialists available on the market. This problem is especially visible in Poland, where the academia is not able to promote enough number of data science specialists and provide them to the industry.

2 The ATENA project

The ATENA (Artificial inTElligence traiNing progrAmme) project is the educational initiative held at Bydgoszcz University of Science and Technology in

Bydgoszcz, Poland as a part of the SPINAKER programme. Thanks to the support from National Agency for Academic Exchange it was possible to create the innovative, short and intensive course for foreign students. The main goal of each of three editions of the ATENA is to gather students, scientists and professionalists, and to provide them the space to share knowledge, discuss and create some powerful, realistic AI-based solutions. The key point of the project is to use numerous modern learning techniques. We worked with students in a blended form: there were both online and on site participants. Lectures were as well remote and in person. We used also the online resources bank provided for self-learning. The project management was performed mostly online due to the SARS-CoV-2 pandemic.

3 Student's projects

3.1 Predictions of biomedical properties based on low-coherence measurements

The main goal of this project was to predict if given sample is healthy or has cervical cancer. In this approach optoelectronic sensor was connected with machine learning algorithm to achieve this goal. Data was gathered by the optical measurements of refractive index of each sample as described in [4]. The data was pre-processed and enriched in following steps:

- Find local and global maximas,
- Filter data by threshold (0.05),
- Determine distance between local maximas,
- Get min/max wavelength,
- Amplitude of 'Amplitude' column,
- Calculate max/min/average/median of local maximas distances,
- Determine y value - if refractive index > 1.3 the tissue is infected otherwise is healthy,
- Get cavity length.

There were several models trained and evaluated: kNN, SVM, Decision Tree Classifier, Gaussian Naïve Bayes, Random Forest Classifier and XGBoost. The best performance on the dataset was achieved by XGBoost model. It provided the results reaching 98% accuracy, 98% precision, 100% recall and 99% F1.

3.2 Prediction the risk of defaulting a loan based on the customer behavior

The aim of the project was to implement a machine learning based methods to detect potential customers who may have problems with repaying a loan taken from a bank. In the research we used a publicly available dataset of people taking loans from banks provided by Kaggle. This collection consist of 280,000 rows and 13 columns, where each column, except last one, contains customer data.

The last column contains a label if customer paid back the loan. Before using machine learning algorithms, data had to be properly prepared. LabelEncoding was used for the column containing information about marital status and car ownership. On the other hand, OneHotEncoding was used for the city, state, house ownership and occupation columns. In order to implement the ML-based methods we divide the dataset as follows: 85% of the data was used for training and 15% for testing models. For the classification stage, three methods were used and compared: Decision Tree (DT), Random Forest (RF) and MLP classifier (MLP). Decision Tree and Random Forest achieved similar results at the score of $F1 = 77\%$ for DT and $F1 = 79\%$ for RF. The MLP provided the less promising results ($F1 = 53\%$).

3.3 Network intrusion detection

The aim of the project was to implement a machine learning-based methods in order to detect malicious activity on the Internet traffic data. In this research the dataset provided by UNB (University of New Brunswick's) was used. It is a collection of internet traffic in some days of the week. It is composed by 6 Million rows and 78 columns. Each column, apart the last one, represents a feature. The last one represents the label. In order to use the ML-based algorithms the pre-processing was needed: casting values (changing every value into numeric format) and handling NaN (the dataset was enough big to remove the records with NaN values). During the research 80% of samples were used for training and 20% for testing. For the classification step two methods were used and compared: Random Forest (RF) [2] and Artificial Neural Network (ANN) [3]. The ANN outperformed reaching the results reaching F1-score 88% - 99% for various classes, when the RF provided the F1-score reaching 67% - 93%.

3.4 Automation of document scanning

The aim of this project was the implementation of an automation mechanism for the recognition of documents. In this case the proposed method was used with sow cards from animal farms that have been previously filled manually and photographed. These documents need to be cropped and straightened, and later, significant data can be extracted from tables. The dataset provided by a local animal farms was very limited (only 17 photos taken in color with different illumination and angles conditions). Thus, the augmentation was needed: flipping, mirroring, brightness modification and shifting. The biggest challenge in the project was to perform the segmentation. Firstly, the masks were painted manually for the whole small 17-elements dataset. Then, the U-net architecture, which is considered as very powerful in such applications [5], was used with satisfying results. It was possible to obtain the accuracy reaching 95%.

4 Conclusions

As presented above, the AI methods can be implemented for solving numerous problems. The propositions of the students' solutions were created during an one week long stay in Poland. Thus, in the future they can be improved and investigated in details. However, the project was an opportunity of students to cooperate in the international teams, to discover some new AI-based architectures and to make the first approach in designing the scientific experiment and run the research.

The AI is an extremely important issue in the computer science studies, but it is often underappreciated in the higher studies programs. ATENA project shows that data science and machine learning can be very interesting and involving for students.

Acknowledgements

This project has received founding from the European Union's PO WER project. It is supported by the Polish National Agency for Academic Exchange in the SPINAKER programme. The total EU funding is 369 490,00 PLN.

References

1. Coordinated plan on artificial intelligence (2018), https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52018DC0795&rid=3
2. Ajdani, M., Ghaffary, H.: Introduced a new method for enhancement of intrusion detection with random forest and pso algorithm. Security and Privacy **4**(2), e147 (2021)
3. Choraś, M., Pawlicki, M.: Intrusion detection approach based on optimised artificial neural network. Neurocomputing **452**, 705–715 (2021)
4. Kruczkowski, M., Drabik-Kruczkowska, A., Marciniak, A., Tarczewska, M., Kosowska, M., Szczerska, M.: Machine learning for predictions of cervical cancer identification–preliminary investigation based on refractive index. preprint available at ResearchSquare (2021)
5. Siddique, N., Paheding, S., Elkin, C.P., Devabhaktuni, V.: U-net and its variants for medical image segmentation: A review of theory and applications. IEEE Access **9**, 82031–82057 (2021)