AGH

AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

Wydział Elektrotechniki, Automatyki, Informatyki i Inżynierii Biomedycznej

Master Thesis

Development of tools for data quality control for ALICE experiment at CERN, using machine learning methods.

Rozwój narzędzi do monitorowania jakości danych, dla eksperymentu ALICE w CERN, z użyciem metod uczenia maszynowego.

Author:	*Bartłomiej Cerek*
Field of Study:	Computer Science
Thesis Supervisor:	*dr hab. Adrian Horzyk, prof. AGH*

Kraków, 2020

Uprzedzony o odpowiedzialności karnej na podstawie art. 115 ust. 1 i 2 ustawy z dnia 4 lutego 1994 r. o prawie autorskim i prawach pokrewnych (t.j. Dz.U. z 2006 r. Nr 90, poz. 631 z późn. zm.): „ Kto przywłaszcza sobie autorstwo albo wprowadza w błąd co do autorstwa całości lub części cudzego utworu albo artystycznego wykonania, podlega grzywnie, karze ograniczenia wolności albo pozbawienia wolności do lat 3. Tej samej karze podlega, kto rozpowszechnia bez podania nazwiska lub pseudonimu twórcy cudzy utwór w wersji oryginalnej albo w postaci opracowania, artystyczne wykonanie albo publicznie zniekształca taki utwór, artystyczne wykonanie, fonogram, wideogram lub nadanie.", a także uprzedzony o odpowiedzialności dyscyplinarnej na podstawie art. 211 ust. 1 ustawy z dnia 27 lipca 2005 r. Prawo o szkolnictwie wyższym (t.j. Dz. U. z 2012 r. poz. 572, z późn. zm.) „Za naruszenie przepisów obowiązujących w uczelni oraz za czyny uchybiające godności studenta student ponosi odpowiedzialność dyscyplinarną przed komisją dyscyplinarną albo przed sądem koleżeńskim samorządu studenckiego, zwanym dalej „sądem koleżeńskim", oświadczam, że niniejszą pracę dyplomową wykonałem(-am) osobiście i samodzielnie i że nie korzystałem(-am) ze źródeł innych niż wymienione w pracy.

....

I would like to express my gratitude to my thesis supervisor, for all the guidance as well as to Professor Jacek Otwinowski and Sebastian Bysiak for valuable advice.

Table of Content

1. Introduction

In recent years, **Machine Learning (ML)** and **Deep Learning (DL)** are being frequently listed among the most prominent technologies. “We are now solving problems with machine learning and artificial intelligence that were… in the realm of science fiction for the last several decades. And natural language understanding, machine vision problems, it really is an amazing renaissance.” This is a quote from Jeff Bezos, CEO of Amazon, one of the biggest companies in the world, that succeeded among others, because of the great usage of data. [1] This is not a singular case, in more and more industries data science is being used to create business leverage and outrun the competition.

A natural question to put forward is, can those technologies be used to also help researchers? The development of new machine learning tools and models is a primarily academic field but its findings were mostly applied in business. Now, also scientists are more willing to use techniques of **Artificial Intelligence (AI)** in their work. Some experiments in medicine, physics or chemistry produce lots of data and automated algorithms are a convenient way to find patterns and dependencies that are difficult to be spotted by humans. Research facilities like CERN employ ML and DL to push forward the limitations of our knowledge. [2] Organized in 2014 The Higgs Machine Learning Challenge brought many AI enthusiasts to work on solving High Energy Physics problems and encouraged the further application of ML in science. [3] This enthusiasm is generally growing as tools and technologies connected to intelligent algorithms are more available and easier to use. Toolkits and libraries like Scikit-learn for Python are getting a vast number of users across many specializations, thanks to the ease of use and universality of application. [4]

There exist, however, problems and limitations in use of machine learning that should be mentioned. Nowadays many scientists discuss the so-called ‘reproducibility crisis’, a growing problem in metascience which is the inability to replicate or reproduce scientific studies. [5] Because ML algorithms are designed to find patterns in given data, even when there are none. They can fixate on noise and cannot asses their own uncertainty. As most of AI models work as black boxes, they take certain inputs and produce outputs, without explaining their decisions. This can be not only deceptive but sometimes even counterproductive when the ultimate goal is understanding the underlying process. Lack of decision argumentation reduces trust in ML methods and can even make it unusable in fields like medicine, where the confidence in decisions is crucial.

Those issues challenge machine learning and deep learning techniques used for scientific research. There is a growing movement of ‘explainable artificial intelligence’. New algorithms are created, that not only make a decision, but also try to argument it. Tools for explaining choices of existing black-box models are also researched and developed. [6][7]

The goal of this thesis is to propose and compare different machine learning and deep learning solutions to quality control task in CERN’s ALICE experiment while addressing the aforementioned issues. The scope of the project includes the analysis of example dataset from ALICE detector, data mining, development of different classification models and proposing pipelines for their usage in unsupervised, supervised and semisupervised cases. Evaluation of the obtained solutions is also provided. An important aspect is the comparison of methods based not only on their accuracy but also on the possibility of practical implementation in the environment of the ALICE experiment.

The thesis contains 8 chapters, bibliography and appendix. First part is dedicated to introduction of quality control task in CERN’s ALICE and description of used in project software tools. Following this, in depth analysis of sample dataset

2. Quality Control in CERN’s ALICE

CERN (Conseil Européen pour la Recherche Nucléaire) is a research organization focused on particle physics. It was established in 1954 and is placed on the Swiss/French border, near Geneva. CERN is well known for its particle accelerators used for high-energy physics research and also for advancing engineering and computer science. [1] W and Z bosons (1983), as well as Higgs boson (2012), were discovered there. [2] Moreover, it is the birthplace of the World Wide Web (www).

LHC (Large Hadron Collider) is the largest particle collider in the world, that was designed and built at CERN and started operating in 2008. It has 27 km in circumference and currently is able to reach the energy of 6.5 TeV per beam. It was constructed to answer some of the fundamental, open questions in physics, most notably regarding the Standard Model and mass of elementary particles. [3]

On LHC’s intersection points are located CERN’s experiments. Biggest of them, ATLAS and CMS are general-purpose particle detectors that allowed for the discovery of Higgs boson. Others are dedicated to specialized research. For the purpose of this thesis, **ALICE** (A Large Ion Collider Experiment) is crucial. Its goal is to study the collisions of heavy ions (like plomb) for the understanding of QCD, the strong-interaction part of the Standard Model. Beyond that, it can take proton beams data at the highest LHC energies to collect references for the heavy-ion program. ALICE weighs 10 000 tonnes and is located underground, close to the village of Saint Genis-Pouilly in France. [4] [5]

Bibliography

TO BE REWRITTEN BY STANDARDS

[1]https://www.cnbc.com/2017/05/08/amazon-jeff-bezos-artificial-intelligence-ai-golden-age.html

[2] https://journals.aps.org/prl/pdf/10.1103/PhysRevLett.121.241803

[3] https://iopscience.iop.org/article/10.1088/1742-6596/664/7/072015/pdf

[4] https://scikit-learn.org/stable/

[5] https://en.wikipedia.org/wiki/Replication_crisis

[7] https://arxiv.org/pdf/1708.08296.pdf).

[8] https://arxiv.org/abs/1602.04938