2024

Artificial intelligence with laws and causality study

Our group operated with two people in this year, one of them (Marcel T. Kurbucz) left the Wigner RCP since then, he continues his scientific work in one of the most prominent institutions, the UCL university in London. We also cooperated with others in order to achieve our scientific goals.

One topic, that we pursued in this year is the continuation of the law-based time series analysis. We published a paper on the methodology of our ideas (c.f. [1]). The main difference between our and the usual supervised learning approaches is that we identify the relevant features by their persistency. If the data for a longer period satisfy a certain rule (law), then it is worth to recognize it as a relevant feature, since they are a good way to characterize the data for an extended period of time. The most simple functional space that can be used to look for a meaningful conserved quantity, is the space of linear functions. This is, actually, the way, how the DNNs, or the brain works. With several linear laws a nonlinear function can also be approximated, like the local straight lines can approximate a function with changing slope.

With this method, we also studied several databases last year, and collected our results for benchmarking purposes. This work is in preprint state as yet (c.f. [2]), waiting the reference process. In this paper we applied an improvement, by using adaptive methods to determine the hyperparameters of our method.

In the last year, we continued a different line of thought as well: this is the causality study. The basic problem with Artificial Intelligence is that it must pre-choose those information, that shall be combined for a meaningful result. For example, in the convolution neural networks (CNN) we assume that local information are correlated, thus a local window can grab the essence of the incoming information. It is also known, that understanding which parts of the incoming data are interdependent results in a more powerful data analysis tool, than just allowing all type of connections to appear, trusting the system that it will cancel those that are not important.

In case of large cross sectional time series, like the ones in the stock market, or the ones coming from a real brain observation, it is far from being evident, which data are worth to combine. As mentioned above, allowing too lot of combinations with trainable weights may result in a poorly performing, or even an incompetent network.

In these cases other, non-AI methods can come in handy. Causality analysis tries to reveal, which data influence each other, without suggesting an actual model of the interdependence. Therefore it is especially good method to select those data elements, from which a good, predictive model can be built.

In the causality analysis we started with Markov chains, and published a paper about the hidden variable exploration [3]. The observation was that the autocorrelation function of the data is in intimate relation with the number of effective states in the Markov chain. More specifically, the autocorrelation function satisfies a linear law, which is related to the characteristic polynomial of the transfer matrix.

We continued our study with the continuous case [4]. In the continuous situation the main problem is that causality analysis methods exist for deterministic cases (based on Taken's theorem), or stochastic cases, but so far there were no viable methods that would be applicable for both of them. We worked out such a unified method, based on the definition of degrees of freedom. The point is that by fixing a certain number of initial conditions, the complete time series is fixed, so new conditions will not influence the distribution of the values in the series. This thought can be generalized to the stochastic case. Practically we can not fix exact values, since we probably do not have enough data that supports a very narrow restriction. Instead, following the ideas of physics, we apply loose conditions that are getting stricter and stricter gradually. In the "continuum limit" we arrive at the strictest restrictions.

If we know the amount of information necessary to maximally determine the observed time series, then we can argue that if X→ Y, (X drives Y), all information showing up in X shows up in Y too, but not vice versa. In this way we can determine various types of causal relations between the different data series.

An independent, but very interesting study was published in [5]. It discusses the relation between the Gini index used to measure social inequalities, and the non-additive entropy.

[1] LLT: An R package for linear law-based feature space transformation, MT Kurbucz, P Pósfay, A Jakovác, SoftwareX 25, 101623

[2] Adaptive Law-Based Transformation (ALT): A Lightweight Feature Representation for Time Series Classification, MT Kurbucz, B Hajós, BP Halmos, VÁ Molnár, A Jakovác, arXiv preprint arXiv:2501.09217

[3] State space reconstruction of Markov chains via autocorrelation structure, A Jakovác, MT Kurbucz, A Telcs, Journal of Physics A: Mathematical and Theoretical 57 (31), 315701

[4] Unified Causality Analysis Based on the Degrees of Freedom, A Telcs, MT Kurbucz, A Jakovác, arXiv preprint arXiv:2410.19469

[5] Analogies and Relations between Non-Additive Entropy Formulas and Gintropy, TS Biró, A Telcs, A Jakovác, Entropy 26 (3), 185

2019

Machine learning and data analytics platform for infectious disease genetics. — Our group's focus is to foster research in data and computation intensive research areas. The last two decades have seen an unprecedented change in almost all areas of sciences. Before that most disciplines were determined by the scarcity of experimental data. The exponential pace of microelectronics development has changed this, on one hand by making available high throughput sensors and digital instruments and on the other by providing high speed computers with large storage and fast interconnecting network. Beyond the almost limitless opportunities there are demanding challenges, too: how to handle the data avalanche from experiments, how to get out the most from information technology in various scientific disciplines, and how to understand and manages the ever-growing complexity of the computational system itself. We study computer networks and systems like it was a "natural phenomena" and with continuously following the technologies, we use them for analyzing science data in various fields from genomics to cosmology.

We are part of a large European H2020 project, COMPARE in which bioinformatics tools are developed for outbreak detection. The health of humans and animals around the world is increasingly under threat due to new and recurring epidemics and foodborne disease outbreaks, which place pressure on health services and the production of livestock. These epidemics also reduce consumer confidence in food and negatively impact trade and food security. The longer it takes from the start of an outbreak of for example Ebola, influenza or salmonella until it is detected and stopped, the greater the consequences. The most important factor in being able to limit the consequences and costs of such outbreaks is the ability to quickly identify the disease-causing microorganisms that are causing the disease. Also, there is the need for knowledge about the mechanisms that cause the disease, and how the bacteria are transmitted to and between humans. The goal of the COMPARE project is a better surveillance system for infectious diseases, to speed up the detection of and response to disease outbreaks among humans and animals worldwide using new genome technology. Our group is responsible for the advanced database and data analysis system which will store, analyse and share the genomic data collected by researchers all over the world. We develop a "virtual research environment", where interested partners can log in, and use the already installed tools, software and data together with their own to do research (Fig. 1). Wigner Cloud is used as a hardware backend for developing the portal. We are also involved in the development of machine learning methods, like artificial neural networks for inferring antibiotic resistance based on the genetic sequences of bacteria.

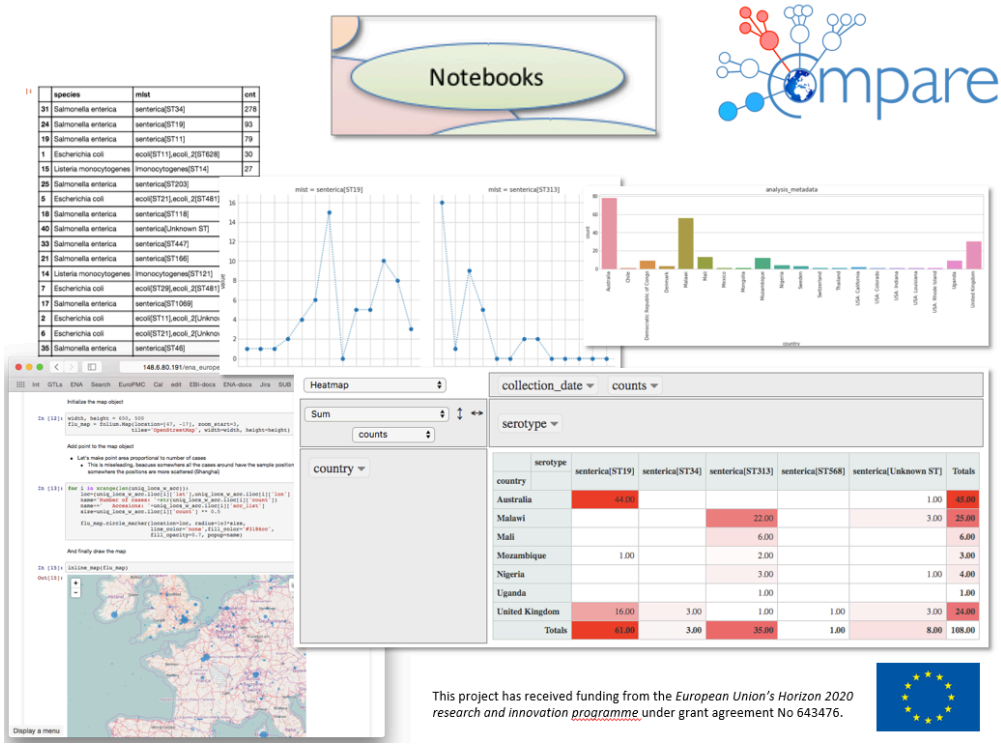

Figure 1. Snapshot of pathogen genome data analysis in the COMPARE Data Hub.

Év
sze, 01/02/2019 - 12:00