Empirical Approach to Machine Learning

Prof Plamen Angelov, PhD, DSc, FIEEE, FIET

School of Computing and Communications, Lancaster University, UK
e-mail: p.angelov@lancaster.ac.uk; www.lancs.ac.uk/staff/angelov

The staggering proliferation of heterogeneous, large scale data sets and streams is recognised as an untapped resource which offers new opportunities for extracting aggregated information to inform decision-making in policy and commerce. However, currently existing methods and techniques for data mining involve a lot of prior assumptions, handcrafting and a range of other *bottleneck issues*: i) scalability – vast amounts of data which require high throughput automated methods (e.g. manual labelling of data samples can be prohibitive); ii) complex, heterogeneous data (including signals, images, text that may be uncertain and unstructured); iii) dynamically evolving, non-stationary data patterns, and the shortcomings of the "standard" assumptions about data distributions; iv) the need to hand craft features, parameters or set thresholds. As a result, a large proportion of the available data remains untapped. The **key challenge** now is to manage, process and **gain *value* and *understanding*** from the vast quantity of **heterogeneous** data **without handcrafting and prior assumptions, at an industrial scale**.

In this talk a newly emerging theoretical framework which we call Empirical Data Analytics will be introduced and described and its relation to the probability, density, centrality, etc. Traditional disciplines of Machine Learning, Data Mining, Pattern Recognition, System Modelling and Identification are well developed. However, current tools often require a number of restrictive assumptions, or handcrafting/manual selection of features, distribution types, parameters, thresholds, etc. Existing algorithms are usually *iterative*, including *internal cycles*. In traditional statistical approaches, averages play a more important role than the individual specifics. Even rapidly emerging AI and computational intelligence approaches require *ad hoc* assumptions and *a priori* decisions (e.g. network depth/ architecture, membership function type and parameters). Furthermore, most existing algorithms assume fixed model structures. This hampers their application to *dynamically evolving* non-stationary data streams and dealing with *shifts* and *drifts*. For example, in cybersecurity, adversaries are often adaptive and intelligent; they exploit the vulnerabilities of traditional systems that are based on fixed prior assumptions, designed for stationary data streams and data generated by the same distribution. Attacks on spam filtering may, for exmaple, include spam messages that are obscured by random misspellings of trigger words; similar problems exist for detecting malware and biometric spoofing.

Motivated by the principle of Occam's Razor [3], we suggest a complete departure from traditional approaches to large-scale data analysis: we advocate recognising the central importance and complexity of real-world data. Our aim is to establish a new paradigm for autonomous data analytics that is based on minimal prior assumptions. The guiding principles of this paradigm are that i) we should avoid assumptions about the statistical properties of the data; ii) the burden of human effort should be shifted away from the large amount of raw data to the top of the knowledge pyramid (see Fig. 2); iii) all new methods for data analytics should be scalable.

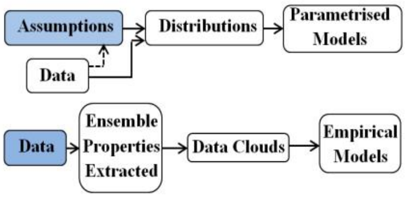

Fig.1 top–traditional approach; bottom–EDA

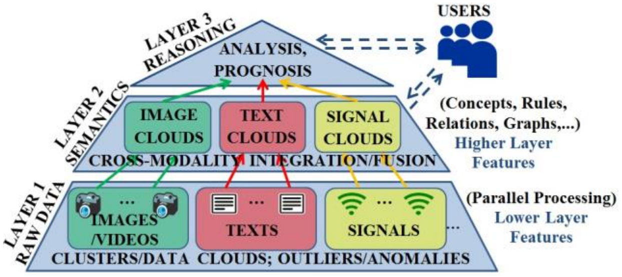

Fig.2 Autonomous Learning Systems within the EDA hierarchical architecture

Within EDA we define *cumulative proximity, typicality, eccentricity, local and global, uni and multimodal density*. ***Typicality*** is particularly interesting, because it resembles (but differs from) the probability density function (pdf), information potential and other similar representations related to system state and structure description and has very close links with laws of physics such as gravitation, intensity and inverse square distance.

Based on this approach fast, transparent deep learning classifiers can be build [4]-[5] which are non-parametric, non-iterative, highly precise and self-developing.

In the talk this new concept will be described as well as a number of applications to various problems.

References:

[1] P. Angelov, Autonomous Learning Systems: From Data Streams to Knowledge in Real time, John Willey and Sons, Dec.2012, ISBN: 978-1-1199-5152-0.

[2] P. P. Angelov, X. Gu, and J. Principe, "A generalized methodology for data analysis," *IEEE Trans. Cybern.*, p. DOI: 10.1109/TCYB.2017.2753880, 2017.

[3] H G Gauch, *Scientific Method in Practice*, Cambridge Univ. Press, 2003.

[4] P. P. Angelov and X. Gu, "MICE: Multi-layer multi-model images classifier ensemble," in *IEEE International Conference on Cybernetics*, 2017, pp. 436–443.

[5] P. Angelov and X. Gu, "A cascade of deep learning fuzzy rule-based image classifier and SVM," in *International Conference on Systems, Man and Cybernetics*, 2017, pp. 1–8.

Biographical data of the speaker:

Prof. Angelov (MEng 1989, PhD 1993, DSc 2015) is a Fellow of the IEEE, of the IET and of the HEA. He is Vice President of the International Neural Networks Society (INNS) for Conference and Governor of the Systems, Man and Cybernetics Society of the IEEE. He has **25+ years of professional experience** in high level research and holds a **Personal Chair** in Intelligent Systems at Lancaster University, UK. He leads the Data Science group at the School of Computing and Communications which includes over 20 academics, researchers and PhD students. He has authored or co-authored **300 peer-reviewed publications in leading journals**, peer-reviewed conference proceedings, 6 patents, two research monographs (by Wiley, 2012 and Springer, 2002) **cited over 6280+ times** with an h-index of 38 and i10-index of 111. His single most cited paper has 810 citations. He has an active research portfolio in the area of computational intelligence and machine learning and internationally recognised results into online and evolving learning and algorithms for knowledge extraction in the form of human-intelligible fuzzy rule-based systems. Prof. Angelov leads numerous projects (including several multimillion ones) funded by UK research councils, EU, industry, UK MoD. His research was recognised by 'The Engineer Innovation and Technology 2008 **Special Award**' and '**For outstanding Services**' (2013) **by IEEE** and INNS. He is also the **founding co-Editor-in-Chief** of Springer's journal on *Evolving Systems* and **Associate Editor** of several leading international scientific journals, including ***IEEE Transactions on Fuzzy Systems*** (the IEEE Transactions with the highest impact factor) of the ***IEEE Transactions on Systems, Man and Cybernetics*** as well as of several other journals such as ***Applied Soft Computing, Fuzzy Sets and***

Systems, Soft Computing, etc. He gave over a dozen **plenary and key note talks** at high profile conferences. Prof. Angelov was General co-Chair of a number of high profile conferences including IJCNN2013, Dallas, TX; IJCNN2015, Killarney, Ireland; the inaugural INNS Conference on Big Data, San Francisco; the 2nd INNS Conference on Big Data, Thessaloniki, Greece and a series of annual IEEE Symposia on Evolving and Adaptive Intelligent Systems. Dr Angelov is the **founding Chair** of the Technical Committee on Evolving Intelligent Systems, SMC Society of the IEEE and was previously chairing the Standards Committee of the Computational Intelligent Society of the IEEE (2010-2012). He was also a member of International Program Committee of over 100 international conferences (primarily IEEE). More details can be found at www.lancs.ac.uk/staff/angelov