MAESTRA: Learning from Massive, Incompletely annotated, and Structured Data

Dragi Kocev[1], Sašo Džeroski[1], Ivica Dimitrovski[2], Michelangelo Ceci[3], Tomislav Šmuc[4] and Joao Gama[5]

[1] Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia

[2] Faculty of Computer Science and Engineering, Universitz Ss Cyril and Methodius, Skopje, Macedonia

[3] Dipartimento di Informatica, Università degli Studi di Bari Aldo Moro, Bari, Italy

[4] Division of Electronics, Ruđer Bošković Institute, Zagreb, Croatia

[5] INESC Technology and Science – INESC TEC, Porto, Portugal

The need for machine learning (ML) and data mining (DM) is ever growing due to the increased pervasiveness of data analysis tasks in almost every area of life, including business, science and technology. Not only is the pervasiveness of data analysis tasks increasing, but so is their complexity. We are increasingly often facing predictive modelling tasks involving one or several of the following complexity aspects: (a) structured data as input or output of the prediction process, (b) very large/massive datasets, with many examples and/or many input/output dimensions, where data may be streaming at high rates, (c) incompletely/partially labelled data, and (d)data placed in a spatio-temporal or network context. Each of these is a major challenge to current ML/DM approaches and is the central topic of active research in areas such as structured-output prediction, mining data streams, semi-supervised learning, and mining network data. The simultaneous presence of several of them is a much harder, currently insurmountable, challenge and severely limits the applicability of ML/DM approaches.

The project will develop predictive modelling methods capable of simultaneously addressing several (ultimately all) of the above complexity aspects. In the most complex case, the methods would be able to address massive sets of network data incompletely labelled with structured outputs. We will develop the foundations (basic concepts and notions) for and the methodology (design and implementation of algorithms) of such approaches. We will demonstrate the potential and utility of the methods on showcase problems from a diverse set of application areas (molecular biology, sensor networks, mutimedia, and social networks). Some of these applications, such as relating the composition of microbiota to human health and the design of social media aggregators, have the potential of transformational impact on important aspects of society, such as personalized medicine and social media.