Doctoral School of Information and Biomedical Technologies
Polish Academy of Sciences (TIB PAN)

SUBJECT: Machine learning from incomplete heterogenous data

SUPERVISOR: Jan Mielniczuk, professor, Institute of Computer Science, PAS

DESCRIPTION:

Inference for incomplete data is an important research area in Machine Learning and Statistics due to ubiquity of such data in practice. Also, available data frequently is heterogenous e.g. exhibiting changes of its distribution in time. Although both problems are intensively studied separately, relatively little is known how to proceed in the situation when the data exhibits both characteristics.

The research proposal focuses on particular types of incomplete data such as data with label noise or positive unlabeled data and specific types of heterogeneity such and label or covariate shift. In particular the starting point would be consideration of heterogenous data with noisy labels. The project builds upon extensive research experience concerning inference for Positive Unlabeled data of the supervisor. The problems of out-of-distribution detection and accounting for imbalance in such a setting are also potential lines of research.

Candidate should have M.Sc. in Mathematics, Computer Science or Engineering, be knowledgeable in Machine Learning and Statistics, including both its mathematical and computational aspects, and possess sufficient computing skills to effectively implement and analyze proposed methods. Scientific curiosity and eagerness to learn are essential.

Candidate should contact the author of the proposal before formal submission of documents (jan.mielniczuk@ipipan.waw.pl).

BIBLIOGRAPHY:

[1] Maity et al (2023) Understanding new tasks through the lens of training data via exponential tilting, ICLR 2023

[2] Cannings et al (2020) , Classification with imperfect training labels, Biometrika 107 (2020)

[3] Yong et al (2023), Holistic view of label noise transition matrix in deep learning and beyond, ICLR 2023

[4] Xuefeng Li et al (2021), Provably end-to-end learning without anchor points, ICML 2021

[5] Lipton et al (2018), Detecting and correcting for label shift with black box predictors, ICML 2018