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

SUBJECT: Long-tail learning from incomplete 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, for multi-class classification the data is frequently imbalanced and accounting for this is called long-tail learning. 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 and positive unlabeled data, and will consider modifications of classical long-tail learning methods to such scenario. In particular the starting point would be studying properties of appropriately modified loss functions adjustment and weighting proposed in Menon et al (2020). The project builds upon former extensive research experience concerning inference for Positve Unlabeled data of the supervisor. The problem of out-of-distribution detection in such a setting is also potential line 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 skill 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] Aditya Menon et al (2021), Long-tail learning via logit adjustment, arXiv preprint arXiv:2007.07314

[2] Timothy 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] Guangxin Su et al (2021), Positive-unlabeled learning for imbalanced data, IJCAI 2021