Politecnico di Torino
Dipartimento di Automatica e Informatica

PhD in Computer and Control Engineering
XXXV cycle

Supervisor
Elena Baralis

Machine Learning with Limited Label Availability: algorithms and applications

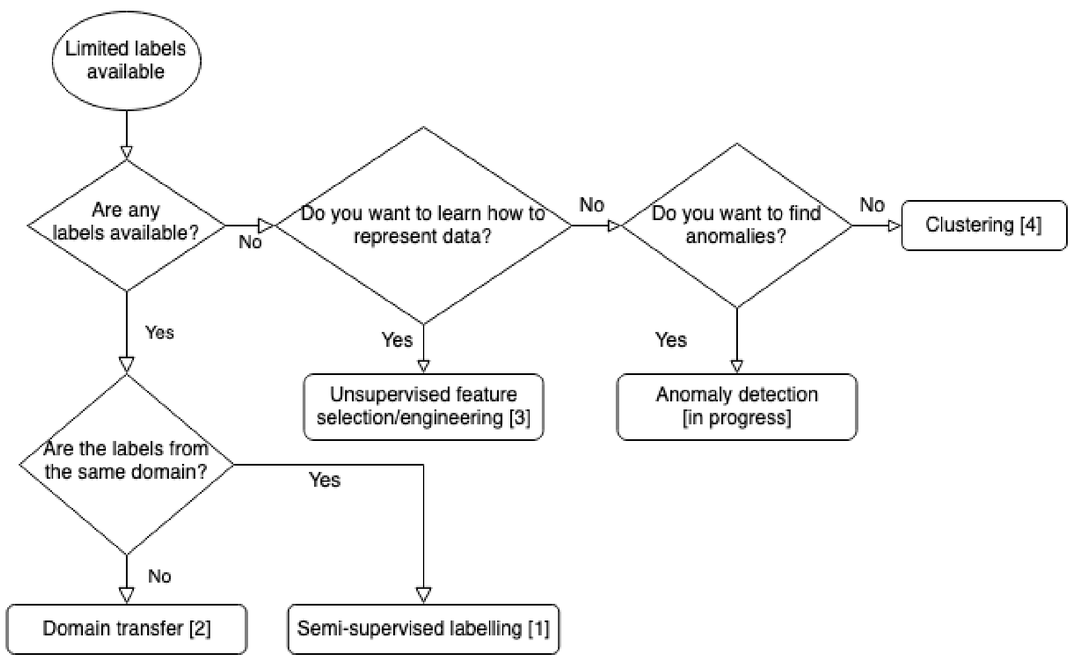
PhD Candidate:
Flavio Giobergia

1. Introduction

The field of machine learning (ML) has thrived for years on the premise that models should be able to learn from large quantities of labelled data. However, the labelling process is often human-based and, as such, expensive. As ML is being more widely adopted, obtaining large quantities of high-quality labels is not a trivial task.

2. Objectives

Our is to explore various scenarios where only limited – if any – labels are available. The scarcity of labels can be mitigated in various ways: either by introducing pseudo-labels, by learning from similar domains or by adopting fully unsupervised techniques. We explored all these aspects in various works, from both an algorithmic and an applied standpoints. The flow chart below shows a non-exhaustive context of applicability of the various approaches.



3. Methods & Results

For scenarios where no labels are available, clustering can be used to characterize points based on similarities found within it. For Self-Organizing Maps, we developed an algorithm that reduced the training time, allowing the adoption of the algorithm to larger datasets and with better quality [4].

For domain transfer in a sentiment analysis setting, we propose learning from high-resource languages (e.g. GB) and transferring to "no-resource" languages [2]. This was done by propagating sentiment information through gradient descent applied on a graph of words in the two languages.

We also approached situations where no clear labels is given, for a predictive maintenance task [1]. We propose obtaining proxy labels through data collected for the specific task. Then we applied a label to each point based on a data distribution, preventing the domain experts from having to label all points manually.

4. Conclusions

Limitations in labels availability hinder the quality of typical ML pipelines. In this work we explored various approaches to reduce the impact of this scarcity. We intend on continuing along this line of research, exploring how few labelled points can be used for the task of anomaly detection.

5. References

1. Giordano, D., Giobergia, F., Pastor, E., La Macchia, A., Cerquitelli, T., Baralis, E., ... & Tricarico, D. (2022). Data-driven strategies for predictive maintenance: Lesson learned from an automotive use case. *Computers in Industry*, *134*, 103554.
2. Giobergia, F., Cagliero, L., Garza, P., & Baralis, E. (2020). Cross-Lingual Propagation of Sentiment Information Based on Bilingual Vector Space Alignment. In *EDBT/ICDT Workshops* (pp. 8-10).
3. Giordano, D., Pastor, E., Giobergia, F., Cerquitelli, T., Baralis, E., Mellia, M., ... & Tricarico, D. (2021). Dissecting a data-driven prognostic pipeline: A powertrain use case. *Expert Systems with Applications*, *180*, 115109.
4. Giobergia, F., & Baralis, E. (2019, December). Fast Self-Organizing Maps Training. In *2019 IEEE International Conference on Big Data (Big Data)* (pp. 2257-2266). IEEE.