Politecnico di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Computer and Control Engineering (*XXXV*cycle)

Machine Learning with Limited Label Availability

Algorithms and Applications

By

Flavio Giobergia

******

Supervisor:
Prof. Elena Baralis

Doctoral Examination Committee:
Prof. Sara Comai, Referee, Politecnico di Milano, Italy
Dr. Dino Ienco, Referee, INRAE, France
Prof. Rosa Meo, Università degli studi di Torino, Italy
Dr. Genoveva Vargas-Solar, CNRS, France
Prof. Silvia Chiusano, Politecnico di Torino, Italy

Politecnico di Torino
2023

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Flavio Giobergia
2023

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

Machine Learning with Limited Label Availability

Flavio Giobergia

An underlying assumption typically made about machine learning algorithms is that the data required for training is both available and curated. However, this assumption does not often hold outside of very specific ad-hoc settings. This in turn hinders the chances of deploying robust machine learning models to address real-world problems, thus reducing the impact that the entire field could actually have.

One of the main obstacles to the successful application of many machine learning algorithms stems from a lack of supervised data: if labels are unavailable, or only available in limited quantities, supervised learning algorithms can struggle during the process of learning useful patterns.

This work of thesis acknowledges the problems that come with limited label availability scenarios and offers contributions within the field. In particular, three main areas of interest are identified, where (i) labelled data is completely unavailable – in which case only unsupervised techniques can be applied, (ii) labelled data is only available in domains other than the one of interest, in which case useful information can still be learned in the label-rich domain and propagated to the label-scarce one and (iii) only reduced amounts of labelled data is available, along with large quantities of unlabelled data.

The thesis addresses these aspects by proposing contributions from both an algorithmic and an applied perspective. In particular, the following contributions that address the aforementioned areas are introduced: (i) improving the training process of existing unsupervised algorithms to reduce their computational complexity, (ii) developing a methodology for propagating representations learned in label-rich domain to label-scarce ones, by means of aligned latent spaces, and (iii) introducing a semi-supervised learning algorithm that learns from self-assigned pseudo-labels based on the introduction of an explicit confidence mechanism.

The proposed contributions are evaluated from both theoretical and empirical perspectives, discussed and contextualized within the existing state of the art. We additionally extensively cover an applied use case, where labels are originally not available but can be inferred in a semi-supervised way.