Titolo: Quantification in the Context of Dataset Shift (QuaDaSh)

Codice MUR: P2022TB5JF

Responsabile scientifico UNIPD: Gian Antonio Susto

Coordinatore nazionale: Alejandro David Moreo Fernandez - Consiglio Nazionale delle Ricerche

Partner-Unità di ricerca: Università degli Studi di Padova - Università di Pisa

CUP: C53D23007970001

Bando: PRIN 2022 PRIN - Decreto Direttoriale n. 1409 del 14-09-2022

Durata: 30/11/2023 - 29/11/2025 (24 mesi)

Budget totale progetto: 225.490,00 €

Budget UNIPD: 91.171,00 €

Abstract del progetto: Quantification is the machine learning task of training estimators of the prevalence values (or "relative frequencies") of the classes in sets of unlabelled data. In principle, this task could be solved by "classify and count": train a classifier, apply it to each unlabelled data item, count how many items have been assigned a given class, and divide by the number of items in the sample. However, research has shown that this approach delivers inaccurate estimates, especially when the unlabelled data exhibit dataset shift (DS), i.e., when the training data and the unlabelled data are not IID samples of the same distribution [Qui09]. Quantification is interesting because most non-idealized application settings do exhibit DS, and because many fields such as the social sciences, political science, market research, and epidemiology, are more interested in quantification than in classification, since they are inherently interested in data at the aggregate level and not at the individual level. Quantification is becoming increasingly important in other disciplines too since, in the era of "big data", we often cannot afford to analyse data at the individual level, and all we can do is analysing them in aggregate form. The goal of the QuaDaSh project is to bring about advances in quantification research and in the understanding of its relation with DS, by:

1. Designing a framework for studying the relations between quantification methods and different types of DS, such as prior probability shift, covariate shift, and concept shift. This will involve (i) the study of experimental protocols that simulate different types of DS, so that the suitability of different quantification algorithms to these different types of DS can be tested, (ii) the study of methods for attaching confidence intervals to the estimations produced by these algorithms, and (iii) the design of new quantification algorithms relying on mixture models and distribution matching.

2. Working on applications of quantification to real-world problems affected by different types of DS, so that these applications can be informed by the above framework and, at the same time, inform it, by feeding back the insights that emerge from the applications. These applications include, among others, the use of quantification for measuring the fairness of machine-learned models, and for computing indicators of poverty, social exclusion, and inequality.

3. Strengthening the quantification ecosystem at the international level, i.e., boosting research in quantification via "horizontal" initiatives that include (i) generating a software library for quantification research that can become a reference for the scientific community, (ii) organizing a data challenge on quantification that allows different quantification methods to be compared in a tightly controlled environment, and (iii) organizing an international workshop on quantification.

Finanziato dall'Unione europea NextGenerationEU | Ministero dell'Università e della Ricerca | Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA