Contribution ID: **39** Type: **not specified**

Methods for Quantifying Dataset Similarity: a Review, Taxonomy and Comparison

Wednesday, 15 May 2024 15:30 (20 minutes)

Quantifying the similarity between datasets has widespread applications in statistics and machine learning. The performance of a predictive model on novel datasets, referred to as generalizability, depends on how similar the training and evaluation datasets are. Exploiting or transferring insights between similar datasets is a key aspect of meta-learning and transfer-learning. In simulation studies, the similarity between distributions of simulated datasets and real datasets, for which the performance of methods is assessed, is crucial. In two- or k-sample testing, it is checked, whether the underlying distributions of two or more datasets coincide.
Extremely many approaches for quantifying dataset similarity have been proposed in the literature. We examine more than 100 methods and provide a taxonomy, classifying them into ten classes, including (i) comparison of cumulative distribution functions, density functions, or characteristic functions; (ii) methods based on multivariate ranks; (iii) discrepancy measures for distributions; (iv) graph-based methods; (v) methods based on inter-point distances; (vi) kernel-based methods; (vii) methods based on binary classification; (viii) distance and similarity measures for datasets; (ix) comparison based on summary statistics; and (x) testing approaches. In an extensive review of these methods the main underlying ideas, formal definitions, and important properties were introduced. The main ideas of the classes are presented here.
We compare the more than 100 methods in terms of their applicability, interpretability, and theoretical properties, in order to provide recommendations for selecting an appropriate dataset similarity measure based on the specific goal of the dataset comparison and on the properties of the datasets at hand. An online tool facilitates the choice of the appropriate dataset similarity measure.

Type of presentation

Contributed Talk

Primary authors: STOLTE, Marieke (Department of Statistics, TU Dortmund University); Dr KAPPENBERG, Franziska (Department of Statistics, TU Dortmund University); Prof. RAHNENFÜHRER, Jörg (Department of Statistics, TU Dortmund University); Dr BOMMERT, Andrea (Department of Statistics, TU Dortmund University)

Presenter: STOLTE, Marieke (Department of Statistics, TU Dortmund University)

Session Classification: Contributed session

Track Classification: Spring Meeting