P-10: A survey of open access databases suitable for machine learning analytics

Päivi Riihimaa[1], PhD
[1]*Digital Health Hub, Department of Medicine, University of Oulu*

Background
Machine learning algorithms require large amounts of high-quality data. Therefore, the scarcity of these remain to be one of the greatest challenges in the data intensive fields of medicine, partly due to the high level of patient data privacy and slow research permission request process. Especially deep learning paradigm requires massive datasets of a size often unattainable in biological studies[1].
This study was undertaken to systematically review open access databases which were qualified as suitable for machine learning analytics with the following criteria:
1) licensing of the dataset allows use for academic, scientific purposes (noncommercial use).
2) dataset is large enough to be analyzed by machine learning. The required sample size is dependent on the purpose of the study, applied algorithm and the number of features to be analyzed, but an estimate of the needed sample size can be estimated using statistical heuristics[2,3].
3) availability of metadata (description of data attributes).

Results
A systematic search was conducted with ten major medical and health dataset collections. The datasets from these collections were evaluated with the abovementioned criteria and described with the following attributes: brief description of data, sample size, number of features, web link and representative example(s) of the published papers based on the dataset.
A list of evaluated datasets will be presented in the poster session, together with the dataset attributes and estimated suitability for clustering, regression, recognition, deep learning or other (e.g. default task). Majority of the found datasets have already been studied in detail using conventional statistical methodology, but some have not been studied with machine learning approach to their full potential. The dataset collection will be available in ResearchGate.

References
1) Camacho, D.M., Collins, K.M., Powers, R.K., Costello, J.C. and Collins, J.J. (2018) Next-Generation Machine Learning for Biological Networks. Cell 173(7): 1581-1592.
2) Figueroa, R.L., Zeng-Treitler, Q., Kandula, S. and Ngo, L.H. (2012) Predicting sample size required for classification performance. BMC Medical Informatics and Decision Making 2012(12):8
3) van der Ploeg, T., Austin, P.C. and Steyerberg, E.W. (2014) Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Medical Research Methodology 2014(14): 137-