The Journal of Open Source Software

Python class defining a machine learning dataset ensuring key-based correspondence and maintaining integrity

Pradeep Reddy Raamana[1] **and Stephen C. Strother**[1, 2]

1 Rotman Research Institute, Baycrest Health Sciences, Toronto, ON, Canada **2** Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada

DOI: 10.21105/joss.00382

Software

- Review
- Repository
- Archive

Summary

A common problem in machine learning is keeping track of the features extracted, and ensuring integrity of the dataset. This is incredibly hard as the number of projects grow, or personnel changes are frequent (hence breaking the chain of hyper-local info about the dataset). This package provides a Python data structure to encapsulate a machine learning dataset with key info greatly suited for neuroimaging applications (or any other domain), where each sample needs to be uniquely identified with a subject ID (or something similar). Key-level correspondence across data, labels (e.g., 1 or 2), classnames (e.g., 'healthy', 'disease') and the related helps maintain data integrity, in addition to offering a way to easily trace back to the sources from where the features have been originally derived.

This data structure also helps ease the machine learning workflow by offering several well-knit methods and useful attributes specifically geared towards neuroscience research.

References

- Raamana, P.R., 2017, *neuropredict: easy, standardized and comprehensive predictive analysis for neuroimaging features*, GitHub. URL: https://github.com/raamana/neuropre