

LADEE Anomaly Detection

Python version: 3.6.1

Python dependencies:

- [scipy](#): miscellaneous statistical functions, machine learning algorithms
- [matplotlib](#): for plotting
- [pandas](#): for reading csv files
- [numpy](#): for math
- [scikit-learn](#): data preprocessing, machine learning algorithms
- [time](#): time and conversions
- [datetime](#): basic date and time types

Files:

- [variables.py](#): this file contains the set of subsystems and the set of variables in each subsystem that we use in anomaly detection.
- [Standardization.py](#) this file imports the set of variables from the data-set and standardizes them [1].
- [Resampling.py](#): this file resamples the features from different subsystems and generates two csv files; one for the beginning of the mission and one for the rest of the mission. In the beginning file, "Resample_Data.csv", the windows are constant time interval. In the second file, "Resample_Data_2.csv" each dark and light period creates a window.
 - [Input parameters](#):
 - [period](#): period of each time window in the beginning of the mission.
 - [number_of_samples_per_window](#): number of sample in each window.
 - [out.csv](#): light and dark periods in the mission
- [Wavelet.py](#): this file extracts the discrete haar wavelet coefficients of each window.
 - [Input parameters](#):
 - [DWT_level](#): the level of wavelet transformation.
- [NonRedundant.py](#): this file removes the redundant features.
 - [Input parameters](#):
 - [threshold](#): the threshold of acceptable correlation between features.
- [Laplacian.py](#): this file selects the features with highest Laplacian score [2].
 - [Input parameters](#):
 - [Number_of_selected_features](#): the number of selected features.

- [Number of Clusters.py](#): this file estimates the number of clusters in the data-set using Calinski and Harabasz's method and Krzanowski and Lai's method [3] and uses hierarchical clustering to generate the clusters.
- [Significant features.py](#): this file uses Biswas et al [4] method to compute significant features of each cluster.
 - [Input parameters](#):
 - [normal_group](#): the normal group
 - [outlier_group](#): the outlier group
 - [threshold](#): this variable represents the minimum required importance that the significant features have to have in distinguishing an outlier from the normal operation.

References:

- [1] Glenn W. Milligan and Martha C. Cooper. A study of standardization of variables in cluster analysis. Journal of classification, 5(2):105–119, 1988.
- [2] He, Xiaofei, Deng Cai, and Partha Niyogi. "Laplacian score for feature selection." In NIPS, vol. 186, p. 189. 2005.
- [3] Yan, Mingjin. "Methods of determining the number of clusters in a data set and a new clustering criterion." PhD diss., Virginia Polytechnic Institute and State University, 2005.
- [4] Biswas, Gautam, Hamed Khorasgani, Gerald Stanje, Abhishek Dubey, Somnath Deb, and Sudipto Ghoshal. "An Approach To Mode and Anomaly Detection with Spacecraft Telemetry Data."