Proposed Implementations/Changes

Ross Brown

March 9, 2022

1 Evaluating

1.1 Current Issue

The algorithms are running on 3 datasets at the moment for an arbitrary number of runs. They are then adjusted/modified manually to visually improve these results which is of little value.

1.2 Solution

Using a set of datasets (currently about 2000 datasets), X, portioned into training, validating, and testing subsets (X_{train}, X_{test}), more robust results can be produced. The idea for these subsets arises from the equivalent idea seen in machine learning.

X_{train} will be used to fit parameters in the algorithm and the X_{test} will have the fitted algorithm applied and the scoring reported. An X_{valid} is not deemed necessary at the moment.

2 Scoring

2.1 Current Issues

Simply taking the mse of the results does not fit in with the biological aspect: finding highly active compounds. The scoring thus needs to fulfil the following criteria:

- Weighting to higher pXa

And would preferably meet the following:

- Lightweight - would rather have computational power used on the active learning than the scoring.
- Independent of dataset size and distribution.

2.2 Solutions

- Weighted mse:

$$\sigma = \sum_i w(y_i - \bar{y})^2$$

  Where w may be:

$$w_i = y_i^\alpha$$

  It is unknown what α would be (so probably 1 according to Occam's razor).

- Number of 'top' results to contain a of the top b true top values.

- Number of iterations needed to get either of the above below a predefined value.

Likely to choose the first solution as simple to implement with a target of 5 iterations. A validation could be used to determine the number of iterations (i.e. look at the rate of improvement and stop at a certain rate).