PR2: Drug Activity Prediction

Published Date:
Oct. 24, 2017, 9:00 a.m.

Deadline Date:
Nov. 7, 2017, 9:00 a.m.

Description:

****
This is an individual assignment.
****

Overview and Assignment Goals:

The objectives of this assignment are the following:

- Use/implement a feature selection/reduction technique.
- Experiment with various classification models.
- Think about dealing with imbalanced data.
- F1 Scoring Metric

Detailed Description:
Develop predictive models that can determine, given a particular compound, whether it is active (1) or not (0).

Drugs are typically small organic molecules that achieve their desired activity by binding to a target site on a receptor. The first step in the discovery of a new drug is usually to identify and isolate the receptor to which it should bind, followed by testing many small molecules for their ability to bind to the target site. This leaves researchers with the task of determining what separates the active (binding) compounds from the inactive (non-binding) ones. Such a determination can then be used in the design of new compounds that not only bind, but also have all the other properties required for a drug (solubility, oral absorption, lack of side effects, appropriate duration of action, toxicity, etc.).

The goal of this competition is to allow you to develop predictive models that can determine, given a particular compound, **whether it is active (1) or not (0).** As such, the goal would be develop the best binary classification model.

A molecule can be represented by several thousands of binary features which represent their topological shapes and other characteristics important for binding.

Since the dataset is imbalanced the scoring function will be the F1-score instead of Accuracy.

Caveats:

+ Remember not all features will be good for predicting activity. Think of feature selection, engineering, reduction (anything that works).

+ The dataset has an imbalanced distribution i.e., within the training set there are only 78 actives (+1) and 722 inactives (0). No information is provided for the test set regarding the distribution.

+ Use the data mining knowledge you have gained until now, wisely, to optimize your results.

Data Description:

The training dataset consists of 800 records and the test dataset consists of 350 records. We provide you with the training class labels and the test labels are held out. The attributes are binary and are presented in a sparse matrix format within train.dat and test.dat. Note that, unlike the CSR matrices we saw before, the values are not listed in the file, since they are always 1.

train.dat: Training set (a sparse binary matrix, patterns in lines, features in columns, with class label 1 or 0 in the first column).

test.dat: Testing set (a sparse binary matrix, patterns in lines, features in columns, no class label provided).

format.dat: A sample submission with 350 entries randomly chosen to be 0 or 1.

Rules:

- This is an individual assignment. Discussion of broad level strategies are allowed but any copying of prediction files and source codes will result in an honor code violation.
- Feel free to use the programming language of your choice for this assignment.
- You are allowed 5 submissions per day.

Deliverables:

- Valid submissions to the Leader Board website: https://coe-cmp.sjsu.edu/clp/ (username is your MySJSU email and your password is your MySJSU password).
- **Canvas Submission of source code and report:**
  - Create a folder called pr2_SJSU-ID
  - Include a 2-page, single-spaced report describing details regarding the steps you followed for feature selection and classifier model development. The report should be in PDF format and the file should be called **report.pdf**. Be sure to include the following in the report:
    1. Name and SJSU ID.
    2. Rank & F1-score for your submission (at the time of writing the report). If you chose not to see the leaderboard, state so.
    3. Your approach.

4. Your methodology of choosing the approach and associated parameters.

- Create a subfolder called src and put all the source code there.
- Archive your parent folder (.zip or tar.gz) and submit via Canvas for PR2.

Grading:

Grading for the Assignment will be split on your implementation (70%), report (20%) and ranking submissions (10%). Extra credit (1% of final grade) will be awarded to the top-3 performing algorithms and to the submission with the most interesting solution (to be judged by Prof. Anastasiu). Note that extra credit throughout the semester will be tallied outside of Canvas and will be added to the final grade at the end of the semester.

Files: In Canvas, you can find

- *Train Data:* train.dat
- *Test Data:* test.dat
- *Format File:* format.dat