COM6912: Individual Dissertation Project:

Machine learning for predicting drug response curves

Marina Krivova

The University of Sheffield, Department of Computer Science, mg.krivova@gmail.com
Supervisor: Dr Mauricio Álvarez

1. Background

Knowledge of a drug sensitivity is crucial in cancer treatment.

Usage of a correct dosage of a correct drug tailored to a particular patient might lower the negative effects of an existing chemotherapy.
But there is inconsistency in methods for estimation of main drug sensitivity parameter - IC50.

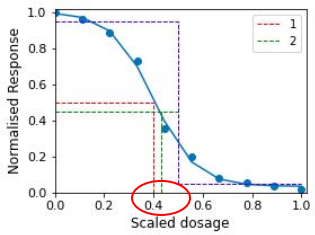

Fig 1. Two methods of evaluation of IC50

Research hypothesis:

1. Is it more efficient to predict drug response curves?
2. Is it better to predict not the separate points but the coefficients of the corresponding fitting function?

Challenges:

1. data quality; **2.** data sparsity; **3.** "fat and short" data

2. Methodology

2.1 Data Preprocessing:

Filtering, Curves fitting, Web-scraping of drug properties

Designed criteria for 3-stage filtration:

1. All the normalised responses should be less than 1
2. The first and last points should form plateaus
3. Specified location of plateaus

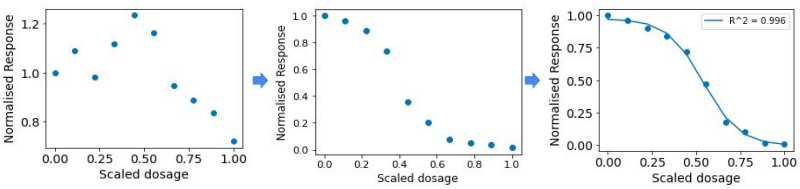

Fig. 2 Illustration of drug curves filtering and fitting

sigmoid_4_param fitting function

$$y = \frac{1}{L + e^{-k*(x-x_0)}} + d$$

Features:

1. Cancer Cell lines features (CCL) - 1073
2. Drug description features - 252
   "Target Pathway" - 23 & "Target" - 229
3. PubChem drug features - 26
4. Max concentration

ΣN **=1352**
Sparsity = 96%

2.2 ML predictions:

"Fit-predict-reconstruct" approach
- Algorithm 1: drug-by-drug training
- Algorithm 2: all-drugs training

"Predict one-by-one" approach - Algorithm 3

10 models:

2 Linear Regressors (Lasso, Ridge)
4 Kernel Ridge (Linear, Sigmoid, RBF, Polynomial)
4 SVR((Linear, Sigmoid, RBF, Poly)
- Feature Importance, - Analysis of Errors

3. Results

3.1. Reconstruction of drug response curves

Approach	Mean R^2	MAE
Aver. Sigmoid	-0.001±0.664	0.280 ±0.086
Algorithm 1	0.786±0.270	0.103 ±0.051
Algorithm 2	0.766±0.355	**0.101 ± 0.054**
Algorithm 3	-	0.200 ±0.023

3.2 Top50 RFE selected features in Algorithm 2

Y	From CCL	From Pub-Chem	From drug targets	From path-way
1	34	5	8	3
2	38	3	6	3
3	29	8	9	4
4	38	3	8	1

4. Main findings

1. Data filtering is necessary
2. Kernel sparse machines are the most effective
3. "Fit-predict-reconstruct" is more accurate
4. Further data reduction is not reasonable
5. Both CCL and drug properties are important
6. Reconstructed curves can be used further

5. Further work:

1. Improvement of predictions
2. Tandem learning, classification tasks