Two-step kernel ridge regression to predict drug-kinase interactions for the IDG-DREAM Drug-Kinase Binding Prediction Challenge

Introduction

Recently, the Research Unit Knowledge-Based Systems (KERMIT) has developed software for a two-step kernel ridge regression method that can be used in a variety of pairwise learning settings. The methods are implemented in the 'xnet' R package and are available via GitHub (https://centerforstatistics-ugent.github.io/xnet/). The IDG-DREAM Drug-Kinase Binding Prediction Challenge represented an ideal opportunity to test this newly developed software package.

Materials and methods

In a first step, the provided raw dataset was processed. Only data that were annotated with 'K DISS', 'LOGKD', '-LOG K', 'KDISS', 'KD', 'LOG K', '-LOG KDISS', '-LOG KD', 'LOG KD', 'Kd' or 'KD' were kept in the dataset. Furthermore, data points having an unspecified compound or target were deleted from the dataset. Secondly, for drug-kinase interactions having multiple measurements, the average of these measurements was computed and other measurements were deleted. In total, this yielded 42730 measured drug-kinase interactions from which machine learning models could learn patterns. More specifically, a two-step kernel ridge regression method was used (Stock *et al.*, 2018). Therefore, kernel matrices were computed that represented the drugs and kinases. For the drugs, molecular fingerprints were calculated based on the provided SMILE representations of the molecules. This was done using RDKit in Python. From these molecular fingerprints, the Tanimoto similarity measure was calculated to construct a kernel matrix. Additionally, a Gaussian interaction profile (GIP) kernel matrix was computed using the Tanimoto similarities (for imputation of missing interaction values) and the scikit-learn toolbox in Python. Both kernel matrices were transformed to a positive semi-definite matrix by iteratively adding small constants to the first diagonal of the matrix. The Tanimoto kernel matrix and GIP kernel matrix were combined to one final drug kernel matrix. Protein kinases were represented using a kernel matrix constructed from pairwise alignment scores. The computed kernel matrices were used as input for a two-step kernel ridge regression model, as was implemented in the 'xnet' R package (https://centerforstatistics-ugent.github.io/xnet/). Using leave-one-out cross validation, the two hyperparameters (one for each kernel matrix) were optimized. Afterwards, a final model was built using these optimized values and was used to make predictions with.

Results and conclusion

Although it was expected for the method and computed kernel matrices to perform well in this setting, results were unsatisfactory. Performances for predictions in round 2 were as follows: an RMSE of 8.466, a rounded spearman correlation of 0.147 and a rounded average AUC of 0.54. In retrospect, computing averages for the interactions occurring multiple times might not have been a good choice. In addition, the transformation to positive semi-definite matrices could have negatively affected the feature representation. In conclusion, although the implemented R package is easy to work with and is useful in a variety of pairwise prediction settings, the method used for this challenge did not yield adequate results. Improvement of the data preprocessing steps and kernel computation steps can improve performance in the future.

Authors contribution statement

Michiel Stock designed the two-step kernel ridge regression methods. Dimitri Boeckaerts processed data for this challenge, implemented scripts for computation of feature representations, used the implemented methods (xnet R package) to train and test a two-step kernel ridge regression model and analysed results as to find the best performing model. Bernard De Baets and Yves Briers provided funding for this project.

References

Stock, M., Pahikkala, T., Airola, A., Waegeman, W. and De Baets, B. Algebraic shortcuts for leave-one-out cross-validation in supervised network inference. Briefings in Bioinformatics. Stock, M., Pahikkala, T., Airola, A., De Baets, B. and Waegeman, W. A comparative study of pairwise learning methods based on kernel ridge regression. Neural Computation 30 (2018), 2245-2283.

Instructions to run the docker (source code under Files in src directory)

```
docker run -it --rm -v ${PWD}/io:/input -v ${PWD}/io:/output
docker.synapse.org/syn18500740/tskrr-dock:9686206
```