BENCHMARKING MULTITASK LEARNING FOR QSARS

Ivan Olier[(1)], Philippa Grace McCabe[(1)], Disha Agarwal[(2)], Sandra Ortega-Martorell[(1)]

(1) Department of Applied Mathematics, Liverpool John Moores University, Liverpool, UK. I.A.OlierCaparroso@ljmu.ac.uk, P.McCabe@2014.ljmu.ac.uk, S.OrtegaMartorell@ljmu.ac.uk

(2) Department of Information and Communication Technology, Manipal Institute of Technology, Manipal, Karnataka, India. Disha.97.agarwal@gmail.com

Keywords: Multitask learning, Random Forest, Quantitative Structure Activity Relationships (QSARs), Drug development.

Abstract. Learning Quantitative Structure Activity Relationships (QSARs) is a key step in drug development. Traditionally, learning small QSAR datasets has been challenging for traditional methods. Recently, multitask learning has been applied to QSAR learning. Our aim with this study is to further understand the capabilities of multitask learning for QSAR learning. Therefore, we applied multitask learning to eighty QSAR datasets, and compared its performance with conventional single-task learning. Our results showed that multitask learning works well especially with the smaller datasets.

1 Scientific Background

Drug development is one of the most important applications of science, as it is an essential part in the treatment of almost all diseases. A key step in drug development is learning Quantitative Structure Activity Relationships (QSARs) [1], [2]. These are functions that predict a compound's bioactivity from its structure. The standard QSAR learning problem is: given a target (usually a protein) and a set of chemical compounds (small molecules) with associated bioactivities (e.g. inhibiting the target), to learn a predictive mapping from molecular representation to activity.

Almost every form of machine learning or statistical methods have been used to learn QSARs. Previous studies have revealed that for large QSAR datasets, Random Forest and Deep Forward Neural Networks are the ones that produce better results [3], [4]. However, small QSAR datasets are challenging for traditional methods. The reasons for this are not fully understood yet, but further to the fact that we are dealing with wide datasets (i.e. datasets with far more predictors than observations), it is also that smaller datasets are likely to be less heterogeneous as they are usually originated from smaller wet lab assays.

Recently, multitask learning has been applied to QSAR learning [5]. Multitask learning is an inductive transfer approach that improves generalization by using the domain information contained in datasets of similar or related tasks. It works by training multiple datasets at once, helping the model learn from multiple sources. It has been argued that multitask learning could be useful for learning small QSAR datasets, which otherwise could not be suitable for standard machine learning algorithms. In [4], it was suggested that multitask learning is beneficial, especially, for smaller QSAR datasets since it acts as a regularizer, overcoming the risk of overfitting.

The aim of this study is twofold: a) to further understand the capabilities of multitask

learning in the very important real-world application area of QSAR learning, and b) to improve the predictive performance of QSAR learning through the use of multitask learning, hoping with that to accelerate the process of drug development while reducing their costs. Motivated by the success of previous studies [4], [5], we are applying multitask learning to eighty QSAR datasets, and comparing its performance with conventional single-task learning.

2 Materials and Methods

2.1 Datasets

The data used in this study was gathered from ChEMBL [6], which is a medicinal chemistry database managed by the European Bioinformatics Institute (EBI). ChEMBL is abstracted and curated from the scientific literature, and covers a significant fraction of the medicinal chemistry corpus. The data consist of information on the drug targets (mainly proteins from a broad set of target families, e.g. kinases), the structures of the tested compounds (from which different chemoinformatic representations may be calculated), and the bioactivities of the compounds on their targets, such as binding constants, pharmacology, and toxicity. We extracted eighty (80) drug targets from ChEMBL with a diverse number of chemical compounds, ranging from 10 to about 6000, with each target resulting in a dataset with as many observations as compounds. The response (output) variable contains the associated bioactivities, which is a real value that ranges from 0 (which indicates no activity) to 12 (the highest possible level of activity). To form the datasets, we calculated the FCFP4 fingerprint representation, which is commonly used in QSAR learning. This representation uses Boolean variables to indicate the presence or absence of a particular molecular substructure in a molecule (e.g. methyl group, benzine ring). The FCFP4 fingerprint implementation generates 1024 such Boolean variables, which are used as predictor variables (features).

2.2 Multitask learning algorithm

Multitask learning has been implemented in several forms. In this study, we opted for the use of an assistant dataset as proposed in [5]. The assistant dataset will row-bind the original (primary) dataset to form an extended version of it. The observations of the assistant dataset are drawn from the other available datasets as a result of the following process:

i. Assuming the primary dataset is split into training and test subsets, for each individual test observation, we calculate the maximum similarity score between this test observation and all the training observations from the same dataset. This maximum similarity score is considered the baseline score or threshold.

ii. Then, we estimate the similarity scores between this test observation and all the available observations from the rest of datasets, and select those observations with higher similarity than the baseline score. The purpose of this step is to choose every molecule in the other datasets having a similarity score with the test observation higher than the threshold.

iii. Previous steps, i and ii, are repeated for every individual test observation.

In order to determine the similarity between two molecules, we used in this study the Tanimoto similarity index as it happens to be one the best choice for fingerprint-based similarity calculation [7].

2.3 QSAR implementation

As mentioned earlier, many statistical and machine learning methods have been applied for QSAR learning, e.g. linear regression, decision trees, neural networks, nearest-neighbor methods, support vector machines, Bayesian networks, relational learning, etc. An exhaustive benchmarking performed previously on several thousands of datasets extracted from the ChEMBL concluded that random forest was typically the best option for QSAR modeling [8]. For this reason, we decided to use this method in the current study for the single and multitask learning experiments. Random forest is an ensemble learning method that can be successfully used for either regression or classification problems. It builds a number of decision trees on bootstrapped training samples [9].

2.4 Performance assessment

The predictive performance of the QSAR learning methods on the datasets was assessed by taking the average root mean squared error (RMSE) with tenfold cross-validation. We were careful to estimate the RMSEs in both, multi and single task learning models, using the same test samples in each individual iteration, hence a fair model comparison could be performed. For each dataset, the performance improvement of using multitask over single-task learning was estimated as follows:

$$improvement = -\frac{RMSE_{single_task} - RMSE_{multitask}}{RMSE_{single_task}} \times 100\% \quad (1)$$

3 Results

Fig. 1 shows comparative assessment in the form of performance improvement of using multitask learning over single-task learning. A positive value indicates that multitask learning has outperformed single task learning. The overall performance improvement, as estimated as the grand average over all datasets, was 26.7%. As seen in the figure, multitask learning beat single task learning 61 times out of 80 datasets. The highest performance improvement was 356.6% (dataset 80 in the figure) whilst the worst was -13.1% (which means that performance actually decayed when using multitask learning).

As reported in [5], multitask learning could be particularly beneficial for tasks with small datasets. We sought any potential correlation between dataset lengths and multitask learning performance. As can be seen in Fig. 2, multitask learning was more likely to produce higher model performance with smaller datasets.

4 Conclusion

The aim with this study was to benchmark multitask learning against the standard single-task learning approach. The QSAR learning problem is an excellent testbed for the machine learning problem of multitask learning. This is because it is well-known that QSAR tasks are tightly related (e.g. target protein similarity).

From the results, we observed that multitask learning can be beneficial for QSAR development. In particular, it is with the smaller datasets that can be taken the most of multitask learning. This is good news for drug development as more data acquisition means more wet lab assays, which are usually costly, time-consuming and hazardous. However, it is also true that dataset length is not the unique factor that drives multitask learning performance (as can be seen in Fig. 2), hence more research is still needed.

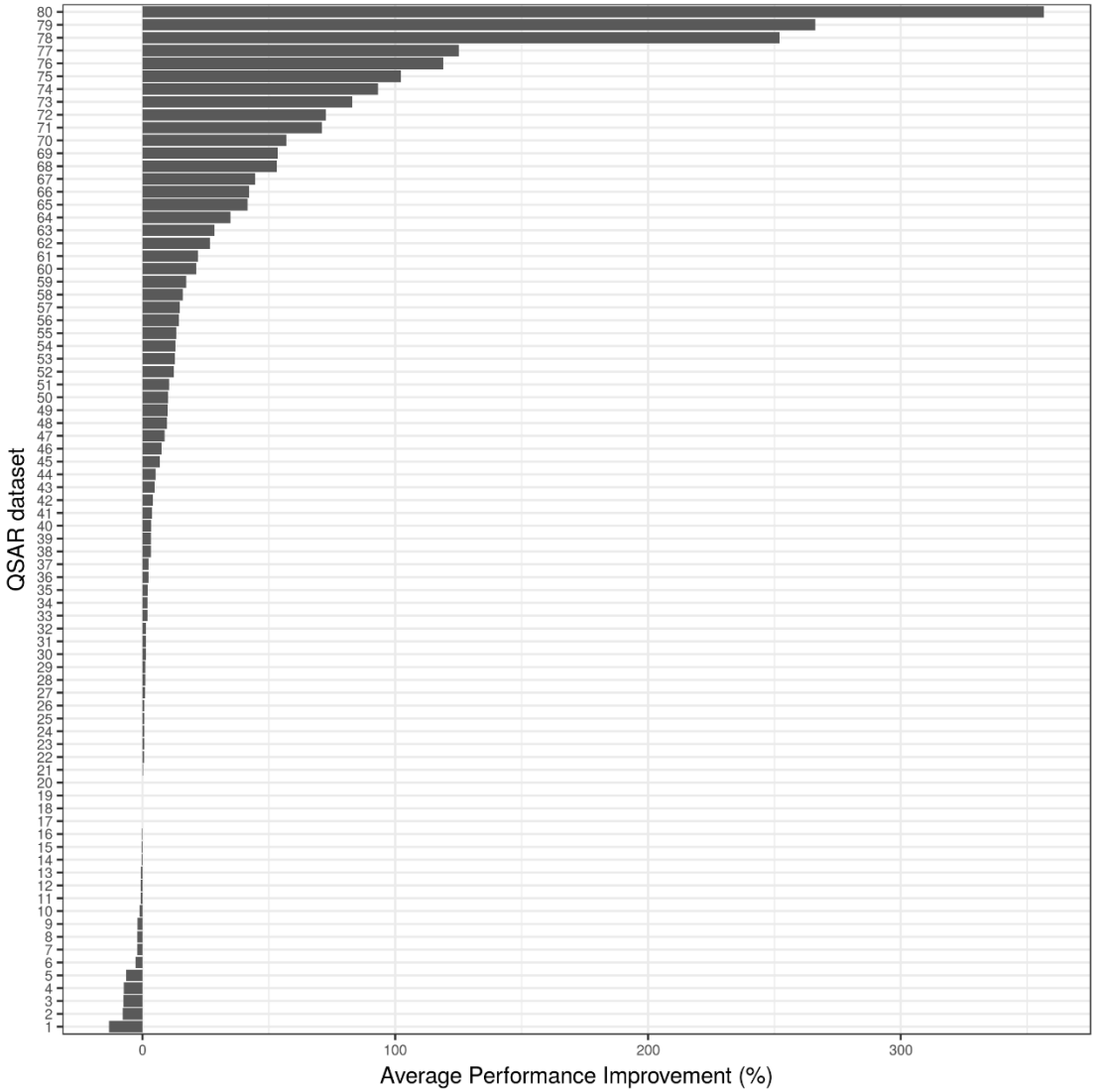


Figure 1: Average model performance improvement of training models using multitask learning over single-task learning. Vertical axis labels identify QSAR datasets.

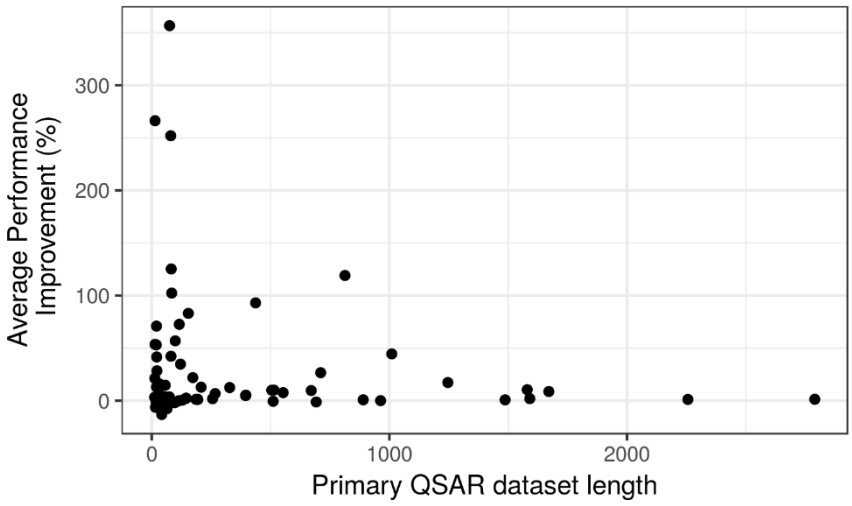


Figure 2: Average performance improvement vs primary QSAR dataset lengths (i.e. original dataset lengths before appending assistant datasets).

References

[1] J. G. Cumming, A. M. Davis, S. Muresan, M. Haeberlein, and H. Chen, “Chemical predictive modelling to improve compound quality,” *Nat. Rev. Drug Discov.*, vol. 12, no. 12, pp. 948–962, Dec. 2013.

[2] A. Cherkasov *et al.*, “QSAR Modeling: Where Have You Been? Where Are You Going To?,” *J. Med. Chem.*, vol. 57, no. 12, pp. 4977–5010, Jun. 2014.

[3] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, “Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling,” *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 6, pp. 1947–1958, Nov. 2003.

[4] J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, and V. Svetnik, “Deep neural nets as a method for quantitative structure-activity relationships,” *J. Chem. Inf. Model.*, vol. 55, no. 2, pp. 263–274, 2015.

[5] Y. Xu, J. Ma, A. Liaw, R. P. Sheridan, and V. Svetnik, “Demystifying Multi-Task Deep Neural Networks for Quantitative Structure-Activity Relationships,” *J. Chem. Inf. Model.*, p. acs.jcim.7b00087, 2017.

[6] A. Gaulton *et al.*, “The ChEMBL database in 2017,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D945–D954, Jan. 2017.

[7] D. Bajusz, A. Rácz, and K. Héberger, “Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?,” *J. Cheminform.*, vol. 7, no. 1, p. 20, Dec. 2015.

[8] I. Olier *et al.*, “Meta-QSAR: a large-scale application of meta-learning to drug design and discovery,” *Mach. Learn.*, vol. 107, no. 1, pp. 285–311, Jan. 2018.

[9] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.