Multiple Task Learning for Quantitative Structure Activity Relationship Learning: Use of a Natural Metric

Noureddin Sadawi[1], Crina Grosan[1], Ivan Olier[2], Larisa Soldatova[1], and Ross D. King[2]

[1] Brunel University, London, UK
[2] The University of Manchester, Manchester, UK

1 Introduction

The task of Quantitative Structure Activity Relationship (QSAR) Learning is to learn a function that inputs the structure of a small molecule (a potential drug) and outputs the predicted activity of the compound against an empirical assay (a test that predicts the potential of the compound being a drug). QSAR learning is a potentially good application area for Multiple Task Learning (MTL) because there is often commonalities in assays. In particular, many assays involve targeting proteins and these proteins are often related. For example, QSAR studies have targeted the proteins dihydrofolate reductase (DHFR) from *Plasmodium falciparum* and *P. vivax* to look for potential anti-malaria drugs. The DHFR from *P. falciparum* is similar, but not identical, to the DHFR from *P. vivax*. It is therefore reasonable to anticipate that it may be better to learn QSARs for both targets at the same time using MTL. It is also noteworthy that the two *Plasmodium* DHFRs are homologous, i.e. they evolved from a common ancestral protein. This enables a natural metric of evolutionary distance to be inferred between the two targets: the closer this distance the more likely the targets are to be similar and MTL to be effective.

In this paper we test two hypotheses: *a*) MTL can improve on standard QSAR learning through use of related targets; *b*) QSAR MTL can be improved by incorporating the evolutionary distance of targets.

2 Methodology

We obtained drug activity data from the publicly available database ChEMBL[3] (as part of our Meta-QSAR project[4]). We collected 454 drug targets, each of which has two or more organisms. To obtain a metric of the similarity of protein targets we first pairwise aligned their sequences and calculated a similarity score based on the alignment (we used *local* alignment).

We performed the following three experiments:

Single Task Learning (STL). We used Random Forest (with 100 trees) as our base learner using the FCFP fingerprint representation of molecules (1024

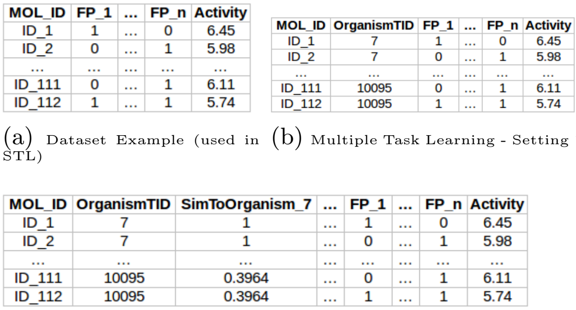
[3] `https://www.ebi.ac.uk/chembl/`

[4] `http://www.meta-qsar.org`

Boolean attributes). We used 10 fold cross-validation to obtain an estimate of the performance for each task (i.e. model).

Multiple Task Learning - Setting 1. In this setting we applied a basic MTL. Our method was to concatenate the datasets for a particular protein target class (e.g. DHFR) and to add an extra *indicator attribute* to each organism (e.g. *P. falciparum*). As shown in Figure 1b, the *OrganismTID* attribute indicates which species the instance came from. We then ran Random Forest (using the same setting as in STL) on the concatenated dataset again using 10 fold cross-validation to obtain an estimate of the performance for each model. Observe that to guarantee contribution from all organisms, we had to ensure that the splits are stratified using this OrganismTID attribute.

Multiple Task Learning - Setting 2. MTL setting 1 makes no use of the similarity between species. We took advantage of this information by adding n extra attributes using the similarity values to the other species as shown in Figure 1c (n is the number of organisms in each drug target). We then ran Random Forest on the big dataset (using the same setting as in STL) and performed the same steps in MTL Setting 1.



MOL_ID	FP_1	...	FP_n	Activity
ID_1	1	...	0	6.45
ID_2	0	...	1	5.98
...	...	...	...	...
ID_111	0	...	1	6.11
ID_112	1	...	1	5.74

(a) Dataset Example (used in STL)

MOL_ID	OrganismTID	FP_1	...	FP_n	Activity
ID_1	7	1	...	0	6.45
ID_2	7	0	...	1	5.98
...	...	...	...	...	...
ID_111	10095	0	...	1	6.11
ID_112	10095	1	...	1	5.74

(b) Multiple Task Learning - Setting 1

MOL_ID	OrganismTID	SimToOrganism_7	...	FP_1	...	FP_n	Activity
ID_1	7	1	...	1	...	0	6.45
ID_2	7	1	...	0	...	1	5.98
...	...	...	...	...	...	...	...
ID_111	10095	0.3964	...	0	...	1	6.11
ID_112	10095	0.3964	...	1	...	1	5.74

(c) Multiple Task Learning - Setting 2

Fig. 1: Datasets for Single and Multiple Task Learning

3 Results & Conclusion

We used Root Mean Squared Error to evaluate performance. To compare performance between settings we counted the number of cases where each setting performs best. As displayed in Figure 2, MTL setting 2 outperforms MTL setting 1 and STL in 611 organisms, the second best setting was MTL setting 1 as it won in 377 occasions and finally STL performed best in 207 occasions. These results are statistically significant.

To conclude, both hypotheses have been confirmed: *a*) MTL can improve on standard QSAR learning; *b*) QSAR MTL can be improved by incorporating the evolutionary distance of targets.

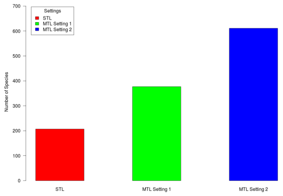


Fig. 2: Experimental Results