Transfer Learning Experiment

December 10, 2014

1 The Idea

- `Target Dataset` is a **small and labeled** dataset from a certain domain
- `Source Dataset(s)` can be one or more **large and labeled** datasets from *related* domains
- If we would like to build a model for the `Target Dataset` using it alone as `Training` dataset, then the model will likely perform undesirably as the dataset is quite small!
- The idea is to make maximum use of the `Source Datasets` to augment the `Target Dataset` and classify data from its domain (this is Instance Transfer Learning)

2 The Data

- The `Source` dataset was from assay TS6 for target DHFR-HS (mCherry). Size = 1435 instances .. Number of Actives = 13
- The `Target` dataset was from assay TS6 for target DHFR-PF (Venus). Size = 1435 instances .. Number of Actives = 27

3 Experiment

- I have `randomly` chosen **48** instances from the target dataset and carried out **10** fold cross validation
- This was repeated **500** times

4 Target Descriptors for Similarity

I have extracted the dipeptide composition which gives a fixed pattern length of 400 (This is the same descriptor I have chosen for our meta-QSAR project). The dipeptide composition encapsulates information about the fraction of amino acids as well as their local order. It is calculated using the following equation:

$$\text{Fraction of dep(i)} = \frac{\text{total number of dep(i)}}{\text{Total number of all possible dipeptides}} \tag{1}$$

Where $dep(i)$ is a dipeptide i out of 400 dipeptides

More information about this descriptor can be found in this paper:
"M. Bhasin, G. P. S. Raghava. Classification of Nuclear Receptors Based on Amino Acid Composition and Dipeptide Composition. Journal of Biological Chemistry, 2004, 279, 23262"

Also, I have understood from reading through several papers that Andreas Bender (Uni of Cambridge) uses it frequently in his experiments. One paper says "these descriptors are simple, yet are known to be quite good at predicting protein structural classes, functional classes and subcellular localizations, Bhasin and Raghava 2004"

5 Target Similarity

Using the descriptor vectors for the two targets (DHFR-HS and DHFR-PF), I have computed the following distance measures:

Distance Type	**Value**
manhattan	1.401771
euclidean	0.1137293
canberra	0.8104851
bray	0.7008855
kulczynski	0.6901639
jaccard	0.8241419
gower	0.46
altGower	0.006581085
horn	0.6733696
mountford	0.9934325
binomial	116.0642
mahalanobis	1.414214

Table 1: Distances between DHFR-HS and DHFR-PF