Greedy and Linear Ensembles of Machine Learning Methods Outperform Single Approaches for QSPR Regression Problems

William Kew & John Mitchell

GUIDE TO SUPPORTING INFORMATION FILES SUPPLIED

“Laura Hughes” Dataset – Data & Scripts

1. method-9-2-14.r

This is the R script used to carry out the computations described in the paper. Note that in the actual production runs a number of classification algorithms were also run, as well as the regression algorithms discussed in the paper. The values of the various tuning parameters can be found in the calls to the relevant routines in this file. The installation of the free statistical software suite R is a prerequisite, see http://www.r-project.org/ for details.

2. folder_names.r

An R script that creates a separate folder to store the results for each different random seed.

3. names-smiles-data-references.xlsx

The names, SMILES one-line structure representations, and experimental logS, logP and melting point data for molecules in the “Laura Hughes” dataset.

4. 262-SMILES-with-descriptors-logS.csv

SMILES and descriptor values used in the calculation of predicted logS values.

5. 262-SMILES-with-descriptors-logP.csv

SMILES and descriptor values used in the calculation of predicted logP values.

6. 262-SMILES-with-descriptors-MP.csv

SMILES and descriptor values used in the calculation of predicted melting points.

7. chemspider_molname_to_smiles_10.t2flow

Taverna workflow for extracting the correct SMILES given a molecule name. This was originally written by Dr Luna De Ferrari and uses the ChemSpider website http://www.chemspider.com/ . The free workflow software Taverna is a prerequisite, see http://www.taverna.org.uk/ for details.

8. histogramlogP.png

9. HistogramofLogS.png

10. HistogramofMP.png

Histograms of the distributions of experimental values of logS, logP and melting point in the “Laura Hughes” dataset.

“Laura Hughes” Dataset – Results

11. CentreScale_logS_R2-Summary.png

12. CentreScale_logS_R2-Summary.xlsx

13. CentreScale_LogS_RMSE-Summary.png

14. CentreScale_logS_RMSE-Summary.xlsx

15. CentreScale_logP_R2 Summary.xlsx

16. CentreScale_logP_R2-Summary.png

17. CentreScale_logP_RMSE Summary.xlsx

18. CentreScale_logP_RMSE-Summary.png

19. CentreScale_MP_R2 Summary.xlsx

20. CentreScale_MP_RMSE Summary.xlsx

These files give spreadsheets and graphs of the results obtained for the “Laura Hughes” dataset with centre scaling. The file names are self-explanatory.

21. PCA_logSR2-Summary.xlsx

22. PCA_logsS_R2-Summary.png

23. PCA_logS_RMSE-Summary.png

24. PCA_logS_RMSE-Summary.xlsx

25. PCA_LogP_R2-Summary.png

26. PCA_logP_RMSE Summary.xlsx

27. PCA_logP_RMSE-Summary.png

28. PCA_logP_R2 Summary.xlsx

29. PCA_MP_R2 Summary.xlsx

30. PCA_MP_R2-Summary.png

31. PCA_MP_RMSE Summary.xlsx

32. PCA_MP_RMSE-Summary.png

These files give spreadsheets and graphs of the results obtained for the “Laura Hughes” dataset with Principal Components Analysis. The file names are self-explanatory.

“Solubility Challenge” Dataset – Data & Scripts

33. method.r

This is the R script used to carry out the computations described in the paper. Note that in the actual production runs a number of classification algorithms were also run, as well as the regression algorithms discussed in the paper. The values of the various tuning parameters can be found in the calls to the relevant routines in this file. The installation of the free statistical software suite R is a prerequisite, see http://www.r-project.org/ for details.

34. SMILES_descriptors.xls

SMILES and descriptor values used in the calculation of predicted logS values.

35. Histogram of LogS.png

Histogram of the distribution of experimental values of logS in the “Solubility Challenge” dataset.

36. CentreScale_r2-summary.csv

37. CentreScale_R2-Summary.png

38. CentreScale_RMSE-Summary.png

39. CentreScale_Summary of Results.xlsx

40. README-centre-scale.txt

These files give spreadsheets and graphs of the results obtained for logS in the “Solubility Challenge” dataset with centre scaling. The file names are self-explanatory.

41. PCA_r2-summary.csv

42. PCA_r2-summary.png

43. PCA_rmse-summary.csv

44. PCA_RMSE-Summary.png

These files give spreadsheets and graphs of the results obtained for logS in the “Solubility Challenge” dataset with Principal Components Analysis. The file names are self-explanatory.

45. Summary of Results.xlsx

A summary of the results obtained for logS in the “Solubility Challenge” dataset.