Table : Average and standard deviations (in brackets) of Area Under Curve (AUC) for different methods applied on the 508 compounds heterogeneous dataset

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Descriptor set used** | | |
| **Combined** | **Basak lab** | **Diudea lab** |
| **PCR** | 0.59 (0.055) | **0.78 (0.038)** | 0.58 (0.057) |
| **PLS** | **0.86 (0.035)** | 0.85 (0.033) | 0.79 (0.038) |
| **Lasso** | 0.72 (0.048) | **0.75 (0.045)** | 0.63 (0.06) |
| **SCAD** | 0.57 (0.061) | 0.58 (0.059) | **0.62 (0.063)** |
| **RF** | **0.81 (0.036)** | 0.80 (0.042) | 0.79 (0.040) |
| **GBM** | 0.80 (0.04) | **0.82 (0.04)** | 0.75 (0.042) |

Table 2: Median and mean absolute deviations (in brackets) of Mean Square Prediction Error (MSPE) for different methods applied on the 95 amines dataset

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Descriptor set used** | | |
| **Combined** | **Basak lab** | **Diudea lab** |
| **PCR** | **29.11 (13.79)** | 57.08 (93.829) | 76.02 (24.72) |
| **PLS** | **18.86 (6.03)** | 19.86 (7.464) | 75.70 (24.689) |
| **Lasso** | **26.85 (9.049)** | 28.72 (8.825) | 72.75 (17.998) |
| **SCAD** | **25.81 (8.962)** | 31.77 (21.442) | 74.94 (18.322) |
| **RF** | **17.25 (6.498)** | 18.98 (6.587) | 84.59 (21.735) |
| **GBM** | **14.79 (5.836)** | 18.03 (6.296) | 74.78 (17.426) |

PCR = Principal Component Regression  
PLS = Partial Least Squares  
Lasso = Least Absolute Shrinkage and Selection Operator  
SCAD = Smoothly Clipped Absolute Deviation penalty  
RF = Random Forest  
GBM = Gradient Boosting Machine

Observations:

For all descriptor sets, PLS has the best performance among all methods, while Boosting performs the best for the 95 amines data.

Methods that depend directly on sparse linear combinations of predictors: Lasso, SCAD do not perform well in either case. This means there is high degree of nonlinearity among the relationship between the responses and predictors, and activities of compounds are more dependent on lower-dimensional subspaces in the predictor space than individual predictors.

PLS performs well in both cases. This implies there are low-dimensional substructures in the predictor spaces that are predictive of the responses.

For the 508 compounds dataset, more predictors do not always equate to better prediction. A reason for this can be the fact that this dataset is composed of chemical compounds from diverse classes. In comparison, the homogeneous 95 compound dataset always gives better prediction with the combined set of predictors than either group of predictors alone.