**Beware of external validation! – A Comparative Study of Several Validation Techniques used in QSAR Modelling**

**Abstract**

**1. Introduction**

**2. Data**

We use two datasets in our study: one simulated and another a well-known chemical activities dataset.

*2.1 Simulated data*

For sample size *n* and number of descriptors *p*, we generate data from the multivariate linear model:

|  |  |
| --- | --- |
|  | **(**1) |

With and being the random error with for = 1, 2, …, *n* and > 0. We fix *n* = 100, and consider three different values of *p*: 100, 500 and 1000. For a fixed p, we first generate rows of the matrix of descriptors as independent and identical draws from a *p*-dimensional normal distribution with mean **0** and covariance matrix . We fix the entries of as

There is often high correlation among chemical descriptors, and when modelling data on hundreds of such descriptors the intrinsic dimensionality of the descriptor data is often much lower than the actual dimension of the predictor space [refs]. We use the above correlation structure to simulate this scenario. For the coefficient vector , we set its first 10 entries as 1 and rest *p* – 10 entries as 0. Finally, we generate elements of by setting , calculate the response variable from (1), and repeat the process for different values of *p*.

*2.2 Congeneric data of 95 amines*

This dataset is due to Debnath *et al* [1]. It contains information on a congeneric set of 95 amine compounds: specifically values on 275 descriptors calculated for each compound, and their mutagenic activities on the *Salmonella typhimurium* strain TA98: as measured by the number of revertants per nmol (in log scale) when a sample compound is applied to a test culture.

< Descriptor information>

**3. Statistical methods**

*3.1 LASSO regression*

For the linear model in (1), the LASSO method proposed by Tibshirani [ref] obtains an estimate of by solving the following minimization problem:

|  |  |
| --- | --- |
|  | (2) |

Where is a tuning parameter. The advantage of using this method is two-fold:

1. Because of the nature of the penalty term the solution is sparse, i.e. some of its entries are exactly set to zero. Thus, LASSO performs simultaneous variable selection and estimation of predictor effects;
2. Unlike linear regression which gives a unique solution only when *n* < *p*, existence and computation of the LASSO solution does not depend on the relative size of *n* and *p*.   
   Thus it is able to tackle high-dimensional regression problems with a large number of predictors but limited sample size (i.e. ).

The large number of descriptors and low intrinsic dimensionality of datasets that are typical of many modern QSAR problems [refs] makes LASSO an ideal candidate for estimation and prediction of chemical activity in such situations.

*3.2 Cross-validation techniques*

We use the following cross-validation techniques to evaluate the predictive capabilities of LASSO models built on the simulated as well as congeneric amine dataset.

***k*-fold cross validation (*k*-fold cv):** We divided the samples randomly into *k* splits, take samples in a split as test set, train a QSAR model on samples outside the test set and predict activity of samples in the test set with that model. Finally we repeat this for all splits to cover all samples.

**Leave-one-out cross validation (LOO-cv):** For a sample of size *n*, we train *n* models, each time taking a distinct sample in the test set to predict the activity of that sample. This can be interpreted as a *n*-fold cross validation.

**External validation:**

**Multiple external validation:**

The tuning parameter for the LASSO regression model in (2) is selected from a range of values using *k*-fold cross-validation. Here we shall take *k* = 5. For this reason, while implementing each of the validation methods mentioned above we need to make sure to incorporate this step every time a model is trained. In this situation, selecting the tuning parameters first on a model built on the full dataset and then predicting for different train-test splits might seem a more intuitive approach. However, this naïve approach uses information from the holdout compounds in the first step, thus providing an inflated estimate of the cross-validated *q*2: which is termed as naïve *q*2 [2].

Thus, we perform cross-validation twice: once to select the best tuning parameter from the training samples, and again to obtain *q*2 values. As an example, for *k*-fold cross-validation the steps for this *two-deep cross validation* procedure will be as follows:

1. Randomly split data into *k* groups.
2. Consider samples in the first split as test set. Select the best tuning parameter by doing a 5-fold CV using the LASSO model in (2) on samples outside the test set.
3. Predict activities of compounds in test set using a LASSO model trained using the best tuning parameter.
4. Repeat steps (b) and (c) considering all other splits as test sets.
5. We now have predictions for all sample compounds. Calculate Prediction Sum of Squares (PRESS) and *q*2 values using these predicted values.

**4. Results**

**5. Discussion**