**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] [1] [2]. 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* [3]. 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.

<Insert table 1>

**Table 1**: Information on descriptor types in the congeneric amines data

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Number of descriptors | Software used | Description |
| TS | 108 |  |  |
| TC | 158 |  |  |
| 3D | 3 |  |  |
| QC | 6 |  |  |

This dataset contains four types of descriptors: topostructural (TS), topochemical (TC), three dimensional (3D) and quantum chemical (QC), in increasing order of computational complexity. Table 1 presents detailed information about these different types of descriptors. There is evidence that while predicting chemical activity through QSAR modelling, the computation-intensive 3D and QC descriptors are largely redundant in presence of a large number of TS and TC descriptors that are computationally easy to calculate [refs] [4] [5] [6]. However, we analyze all four types of descriptors in this paper for the sake of completeness, and because the statistical model used explicitly involves variable selection to automatically filter out variables that are not predictive enough.

**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 [7].

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**

*4.1 Simulated dataset*

For each of the validation methods applied, we report their *q*2 and PRESS obtained using the two-deep method described above.

<Insert table 1>

**Table 2:** Performance of all validation methods on simulated data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *q*2 | | | | |
| Number of predictors (*p*) | | 100 | 500 | 1000 |
| LOO-cv | | 0.80 | 0.74 | 0.71 |
| 5-fold cv | | 0.76 | 0.60 | 0.28 |
| External validation | Min | 0.34 | 0.06 | -0.13 |
| 25th percentile | 0.72 | 0.62 | 0.48 |
| Median | 0.79 | 0.69 | 0.58 |
| 75th percentile | 0.84 | 0.76 | 0.68 |
| max | 0.95 | 0.89 | 0.87 |
| Repeated external validation | | 0.77 | 0.66 | 0.55 |
| PRESS | | | | |
| Number of predictors (*p*) | | 100 | 500 | 1000 |
| LOO-cv | | 1.66 | 2.36 | 2.77 |
| 5-fold cv | | 1.92 | 3.78 | 6.82 |
| External validation | Min | 0.26 | 0.86 | 1.01 |
| 25th percentile | 1.23 | 1.82 | 2.65 |
| Median | 1.63 | 2.48 | 3.35 |
| 75th percentile | 2.04 | 3.12 | 4.86 |
| max | 3.73 | 6.61 | 18.97 |
| Repeated external validation | | 1.70 | 2.59 | 4.24 |

**Table 2** reports values of the two metrics for the four validation techniques, considering the three different number of predictors. For external validation, we report the minimum, 25th percentile, median, 75th percentile and maximum of PRESS and *q*2 from the 100 train-test splits performed during the multiple external validation process. For repeated external validation we report average PRESS and *q*2 over all repetitions. LOO-cv has the best performance across all predictor dimension and both metrics. All methods perform worse as dimension of the descriptor space grows, which is expected because of higher amount of noise introduced by more predictors.

The main issue with external validation, which previous studies (e.g. [refs]) have not captured, is the high degree of variability in its performance depending on which subset of the full data is chosen as the validation sample. The minimum and maximum values indicate that depending on the train-test split, the two-deep *q*2 can vary between 0.34 to 0.95 for *p* = 100, 0.06 to 0.89 for *p* = 500 and 0.01 to 0.87 for *p* = 1000. For *p* = 100, About 50% of the external validation splits have worse performance than LOO-cv for both *q*2 and PRESS, which goes up to around 75% for *p* = 1000. This indicates that for higher number of predictors, LOO-cv is more likely to result a QSAR model that is more predictive. In 2 of the 100 random splits the external validation turned out to be negative. This means that PRESS is more than the total sum of squares in the test set, indicating very high amount of noise in the fitted model, i.e. severe overfitting.

*4.2 Amines dataset*

We report results from the LASSO model validation analysis of the 95 compounds congeneric amines dataset in **Table 3**. In this case, both LOO and 5-fold cv have larger two-deep *q*2 values than repeated external validation, as well as half of the random external validation splits. The minimum *q*2 value for external validation is as low as 0.15. One of the random train-test splits in external validation yielded a *q*2 value of -0.003. This underscores a severe limitation of the external validation procedure: if such a split of a real-world dataset is used to validate a QSAR model, the whole modelling practice becomes nothing but a waste of resources.

<Insert table 2>

**Table 3**: Performance of all validation methods on 95 amines data

|  |  |  |  |
| --- | --- | --- | --- |
| Number of predictors (*p*) | | *q*2 | PRESS |
| LOO-cv | | 0.77 | 0.86 |
| 5-fold cv | | 0.73 | 1.05 |
| External validation | Min | -0.003 | 0.27 |
| 25th percentile | 0.65 | 0.61 |
| Median | 0.73 | 0.88 |
| 75th percentile | 0.83 | 1.28 |
| max | 0.94 | 2.01 |
| Repeated external validation | | 0.71 | 0.97 |

**5. Discussion**

**<Multiple peaks point>**

QSAR modelling is extensively used in academia and industry setup for virtual screening of chemical compounds **[ref]**. These compounds often have lasting impact in human lives and the environment around us. In this situation, a *laissez-faire* use of external validation using small validation sets can have enormous consequences if the wrong compounds get selected in the screening procedure. Thus, it is difficult to overstate the importance of proper, stable and rigorous validation methods.

# **References**

|  |  |
| --- | --- |
| [1] | S. Majumdar and S. C. Basak, "Exploring intrinsic dimensionality of chemical spaces for robust QSAR model development: A comparison of several statistical approaches," *Curr. Comput. Aided Drug Des.,* vol. 12, no. 4, pp. 294-301, 2016. |
| [2] | S. C. Basak, V. R. Magnuson, G. J. Niemi, R. R. Regal and G. D. Veith, "Topological indices: their nature, mutual relatedness, and applications," *Mathematical Modelling,* vol. 8, pp. 300-305, 1987. |
| [3] | A. Debnath, G. Debnath, A. Shusterman and C. Hansch, "A QSAR Investigation of the Role of Hydrophobicity in Regulating Muagenicity in the Ames Test: 1. Mutagenicity of Aromatic and Heteroaromatic Amines in Salmonella typhimurium TA98 and TA100," *Environ. Mol. Mutagen.,* vol. 19, pp. 37-52, 1992. |
| [4] | S. C. Basak, B. D. Gute and G. D. Grunwald, "A hierarchical approach to the development of QSAR models using topological, geometrical and quantum chemical parameters," in *Topological Indices and Related Descriptors in QSAR and QSPR*, J. Devillers and A. T. Balaban, Eds., Amsterdam, The Netherlands, Gordon and Breach Science Publishers, 1999, pp. 675-696. |
| [5] | S. Majumdar and S. C. Basak, "Adapting interrelated two-way clustering method for quantitative structure-activity relationship (QSAR) modeling of mutagenicity/non-mutagenicity of a diverse set of chemicals," *Curr. Comput. Aided Drug Des.,* vol. 9, pp. 463-471, 2013. |
| [6] | S. Majumdar and S. C. Basak, "Prediction of Mutagenicity of Chemicals from Their Calculated Molecular Descriptors: A Case Study with Structurally Homogeneous versus Diverse Datasets," *Curr. Comput. Aided Drug. Des.,* vol. 11, pp. 117-123, 2015. |
| [7] | D. Hawkins, S. Basak and D. Mills, "QSARs for chemical mutagens from structure: ridge regression fitting and diagnostics," *Environ. Toxicol. Pharmacol.,* vol. 16, pp. 37-44, 2004. |
| [8] | Y. Zuo and R. Serfling, "General notions of statistical depth functions," *Ann. Statist.,* vol. 28, pp. 461-482, 2000. |
| [9] | H. Xu, C. Caramanis and S. Mannor, "Outlier-Robust PCA: The High-Dimensional Case," *IEEE Trans. Inf. Theory,* vol. 59, pp. 546-572, 2013. |
| [10] | H. Wiener, "Structural determination of paraffin boiling points," *J. Amer. Chem. Soc.,* vol. 69, pp. 17-20, 1947. |
| [11] | R. P. Verma and C. Hansch, "An approach toward the problem of outliers in QSAR," *Bio. Med. Chem.,* vol. 13, pp. 4597-4621, 2005. |
| [12] | N. Trinajstić, Chemical Graph Theory, Boca Raton, FL: CRC Press, 1992, p. 352. |
| [13] | R. Todeschini and V. Consonni, Molecular Descriptors for Chemoinformatics, New York, NY: Wiley-VCH, 2009. |
| [14] | J. J. Sylvester, "On an application of the new atomic theory to the graphical representation of the invariants and covariants of binary quantics, with three appendices," *Amer. J. Math.,* vol. 1, pp. 64-125, 1878. |
| [15] | J. Stewart, MOPAC Version 6.00, QCPE #455, Frank J. Seiler Research Laboratory: US Air Force Academy, CO, 1990. |
| [16] | J. V. Soderman, CRC Handbook of Identified Carcinogens and Noncarcinogens: Carcinogenicity-Mutagenicity Database, Boca Raton, FL: CRC Press, 1982. |
| [17] | C. Raychaudhury, S. K. Ray, J. J. Ghosh, A. B. Roy and S. C. Basak, "Discrimination of isomeric structures using information-theoretic topological indices," *J. Comput. Chem.,* vol. 5, pp. 581-588, 1984. |
| [18] | M. Randic, "Characterization of molecular branching," *J. Amer. Chem. Soc.,* vol. 97, pp. 6609-6615, 1975. |
| [19] | R Core Team, *R: A Language and Environment for Statistical Computing version 3.1.1,* 2014. |
| [20] | A. H. M. Nandy and S. C. Basak, "Mathematical descriptors of DNA sequences: Development and application," *Arkivoc,* vol. 9, pp. 211-238, 2006. |
| [21] | S. Majumdar, "Robust estimation of principal components from depth-based multivariate rank covariance matrix," 2015. [Online]. Available: http://arxiv.org/abs/1502.07042. |
| [22] | M. Lajiness, Molecular similarity-based methods for selecting compounds for screening. In Computational Chemical Graph Theory, D. H. Rouvray, Ed., Commack, NY: Nova, 1990. |
| [23] | L. B. Kier and L. Hall, Molecular Structure Description: The Electrotopological State, San Diego, CA: Academic Press, 1999. |
| [24] | L. B. Kier and L. H. Hall, Molecular Connectivity in Chemistry and Drug Research, New York, NY: Academic Press, 1976. |
| [25] | M. Karelson, Molecular Descriptors in QSAR/QSPR, New York, NY: Wiley-Interscience, 2000. |
| [26] | D. Janežič, A. Miličević and S. &. T. N. Nikolić, Graph-Theoretical Matrices in Chemistry, Boca Raton, FL: CRC Press, 2015. |
| [27] | H. Hosoya, "Topological Index. A newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons," *Bull. Chem. Soc. Jpn.,* vol. 44, pp. 2332-2339, 1971. |
| [28] | H. Gonzalez-Diaz and C. R. Munteanu, Eds., Topological Indices for Medicinal Chemistry, Biology, Parasitology, Neurological and Social Networks, New York, NY: Transworld research, 2011. |
| [29] | P. A. Filip, T. S. Balaban and A. T. Balaban, "A new approach for devising local graph invariants: derived topological indices with low degeneracy and good correlation ability," *J. Math. Chem.,* vol. 1, pp. 61-83, 1987. |
| [30] | J. Devillers and A. T. Balaban, Eds., Topological Indices and Related Descriptors in QSAR and QSPR, Amsterdam, Netherlands: Gordon and Breach, 1999, p. 811. |
| [31] | B. M. Brown, "Statistical Use of the Spatial Median," *J. R. Statist. Soc. B,* vol. 45, pp. 25-30, 1983. |
| [32] | D. Bonchev, Information Theoretic Indices for Characterization of Chemical Structures, Chichester, UK: Research studies Press, 1983. |
| [33] | S. Basak, D. Mills, B. Gute, A. Balaban, k. K. Basa and G. Grunwald, "Use of Mathematical Structural Invariants in Analyzing, Combinatorial Libraries: A Case Study with psoralen Derivatives," *Curr. Comput. Aided Drug Des.,* vol. 6, pp. 240-251, 2010. |
| [34] | S. Basak, G. Grunwald and A. Balaban, "TRIPLET," Copyright of the Regents of the University of Minnesota, 1993. |
| [35] | S. Basak, V. R. Magnuson, G. J. Niemi and R. R. Regal, "Determining Structural Similarity of Chemicals using Graph-Theoretic Indices," *Disc. Appl. Math.,* vol. 19, pp. 17-44, 1988. |
| [36] | S. C. Basak, "Use of molecular complexity indices in predictive pharmacology and toxicology: A QSAR approach," *Med. Sci. Res.,* vol. 15, pp. 605-609, 1987. |
| [37] | S. C. Basak, D. Mills, B. D. Gute and D. M. Hawkins, "Predicting Mutagenicity of Congeneric and Diverse Sets of Chemicals Using Computed Molecular Descriptors: A Hierarchical Approach," in *Quantitative structure-activity relationship (QSAR) models of mutagens and carcinogens*, R. Benigni, Ed., Boca Raton, FL, CRC Press, 2007, pp. 215-242. |
| [38] | S. C. Basak, D. K. Harriss and V. R. Magnuson, "POLLY v2.3," Copyright of the University of Minnesota, 1988. |
| [39] | S. C. Basak, "Philosophy of Mathematical Chemistry: A Personal Perspective," *Hyle- Int. J. Phil. Chem.,* vol. 19, pp. 3-17, 2013. |
| [40] | S. C. Basak, "Molecular Similarity and Hazard Assessment of Chemicals: A Comparative Study of Arbitrary and Tailored Similarity Spaces," *J. Eng. Sci. Manage. Educ.,* vol. 7, pp. 178-184, 2014. |
| [41] | S. C. Basak, "Mathematical Structural Descriptors of Molecules and Biomolecules: Background and Applications," in *Advances in Mathematical Chemistry and Applications, volume 1*, S. C. Basak, G. Restrepo and J. L. Villaveces, Eds., Bentham eBooks, Bentham Science Publishers and Elsevier, 2015, pp. 3-23. |
| [42] | S. C. Basak, "Mathematical Descriptors for the Prediction of Property, Bioactivity, and Toxicity of Chemicals from their Structure: A Chemical-Cum-Biochemical Approach," *Curr. Comput. Aided Drug. Des.,* vol. 9, pp. 449-462, 2013. |
| [43] | S. C. Basak, V. R. Magnusson, G. J. Niemi and R. R. Regal, "Determining structural similarity of chemicals using graph-theoretic indices," *Dictrete Appl. Math.,* vol. 19, pp. 17-44, 1988. |
| [44] | S. C. Basak and G. D. Grunwald, "APProbe," Copyright of the University of Minnesota, 1993. |
| [45] | S. C. Basak, G. Restrepo and J. L. Villaveces, Eds., Advances in Mathematical Chemistry and Applications, volume 1 & 2, Bentham e-books, Bentham Science Publishers and Elsevier, 2015. |
| [46] | K. Balasubramanian and S. C. Basak, "Characterization of isospectral graphs using graph invariants and derived orthogonal parameters," *J. Chem. Inf. Comput. Sci.,* vol. 38, pp. 367-373, 1998. |
| [47] | A. T. Balaban, "Distance Connectivity Index," *Chem. Phys. Lett.,* vol. 89, pp. 399-404, 1982. |
| [48] | K. P. Adragni and R. D. Cook, "Sufficient dimension reduction and prediction in regression," *Phil. Trans. R. Soc. A,* vol. 367, pp. 4385-4405, 2009. |
| [49] | Sybyl Version 6.2, St. Louis, MO: Tripos Associates, Inc., 1995. |
| [50] | MolconnZ v4.05, Quincy, MA: Hall Ass. Consult., 2003. |