CAse studies on the Development and Application of in-Silico Techniques for Environmental hazard and Risk assessment

Predictive uncertainty by analogy – does it make sense?

Ullrika Sahlin[1], Nina Jeliazkova[2], Tom Aldenberg[3], Jonna Stålring[4], Tomas Öberg[1]
[1]Linneaus University, Sweden, [2]Ideaconsult Ltd., Bulgaria, [3]RIVM, Netherlands, [4]Safety Assessment, AstraZeneca R&D, Mölndal, Sweden

Case study: pIGC50 Tetrahymena pyriformis

BACKGROUND

Decision makers in e.g. chemical regulation or drug design, desires to be able to evaluate the reliability in a prediction of query compound. The associated predictive uncertainty from a QSAR regression and prediction reliability may be sensitive to under what premises the predictive distribution is assessed [1].

One way to assess the predictive distribution from QSAR regression is to assume a parametric probability distribution (e.g. Gaussian) and assess its predictive standard error sd_{PRESS} from the PRedictive Error Sum of Squares (PRESS) (Eq 1), where n is the size of the training data set and the degrees of freedom (n-p-1) could be determined by letting p be the number of latent variables in a PLS regression.

Predictive uncertainty by analogy

Whereas PRESS generates the same predictive error for every compound predicted by a QSAR, predictive error could alternatively be allowed to vary from compound to compound. Such assessment could be based on analogy saying that "compounds that are similar are predicted with similar predictive error".

Assessing predictive uncertainty by analogy reasoning exists within QSAR modeling, but mostly with the purpose of assess reliability in predictions. In the same way as the analogy not always work for QSARs, predicting uncertainty by analogy needs to be motivated both from a theoretical and empirical point of view.

OBJECTIVE

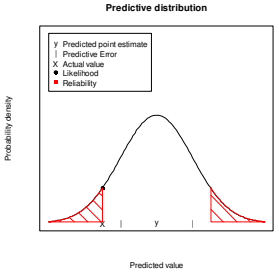
We have challenged the traditional Leave-One-Out estimated PRESS to a weighted PRESS (modified from [2]) for assessing predictive uncertainty.

Weighted PRedicted Error Sum of Squares

We defined wPRESS as a weighted sum of squares (Eq 2). The weight $w_{q,j}$ measures the similarity between the query compound q and the training data. Similarity can e.g. be determined in respect to molecular structure or in respect two how well defined the model is close to the query compound q. The latter being a property of the Applicability Domain (AD) from a statistical point of view (Table 1). The wPRESS can alternatively be defined as a k Nearest Neighbor average (Eq 3).

Figure 1. A probabilistic framework for regression reliability assessment based on the predictive distribution.

$$Y_q \sim N(\hat{y}_q, sd_{q,DPRESS})$$



$$sd_{PRESS} = \sqrt{\frac{PRESS}{(n-p-1)}} = \sqrt{\frac{\sum_{j=1}^{n}(y_j - \hat{y}_{-j})^2}{(n-p-1)}} \quad \text{Eq 1}$$

$$sd_{q,wPRESS} = \sqrt{\frac{wPRESS_q}{(n-p-1)}} \quad \text{Eq 2}$$

$$wPRESS_q = \frac{\sum_{j=1}^{n} w_{q,j} n (y_j - \hat{y}_{-j})^2}{\sum_{j=1}^{n} w_{q,j}}$$

$$wPRESS.kNN_q = \frac{1}{k} \sum_{j \in kNN(w_{q,j})} n(y_j - \hat{y}_{-j})^2 \quad \text{Eq 3}$$

A probabilistic framework for reliability assessment

is based on inference from the predictive distribution of the endpoint of a query compound Yq (Figure 1).

Reliability in the prediction of a single compound is 1- "the confidence level of the smallest confidence interval covering the actual value".

A higher likelihood is, everything else equal, a more reliable assessment of the predictive distribution.

THEORETICAL ARGUMENTS for using analogy to assess predictive uncertainty

The ordinary linear regression model assumes errors to have equal variance (note not the same as predictive error). This assumption can be violated by several reasons. Non-homogeneous variance in regression errors could be an indication of a miss-specified model. Non-homogeneity may arise from unexplained variation by missing important descriptors, descriptor dependent variation or non-linear elements that influence model predictivity. It could also arise when the endpoint variable have not been transformed properly to ensure equal variance in errors. Even non-linear transformations may be necessary.

Unexplained variation may cause non-homogeneous error in predictions, especially under poor conditions of predictive modeling, such as small data set or large associated experimental variability. As a consequence of non-homogeneous variance in errors, may predictivity in less dense areas of the Applicability Domain be lower because the true variability by chance is less described there.

When the assumption of homogeneous variance in errors is true, the predictive variance is expected to be larger in less defined areas of the AD. Predictive variance from Least Squares regression is analytically shown to increase with leverage. i.e. the distance to the center of the AD. The leverage effect could be an argument against PRESS – treating all predictive variances as equal. However, when leverages are small in comparison to the error, e.g. for large data sets, the predictive uncertainty are almost equal over the AD.

Thus, similarity measured with respect to properties of the AD is from a statistical perspective easier to justify for using analogy in the assessment of predictive uncertainty.

Figure 2. Predictive Error based Euclidean distance 5NN.

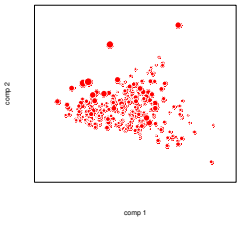


Table 2 pIGC50 Tetrahymena pyriformis

	10 Random partitions	External I	External II
PRESS	4	3	2
Euclidean	2	1	1
Euclidean.5NN	6	6	6
Mahalanobis	3	4	5
Mahalanobis.5NN	7	7	9
Leverage	1	2	3
Leverage.5NN	7	8	8
Density	5	5	4
Density.5NN	9	9	7

Table 1

Similarity measure	Weight in wPRESS*
Similar with respect to Molecular structures	**1/distance**
Pairwise Euclidean distance	
Pairwise Euclidean distance based on Fingerprints	
Similar with respect to an AD property	**1/abs(difference)**
Leverage distance to AD	
Mahalanobis distance to AD	
Probability to be in the AD based on AD density	

*In practice the distances e.d. were 1/n + distance to avoid dividing by zero [2]

CASE STUDY

Euclidean and Leverage based wPRESS were (however not significant) higher ranked than PRESS (Table 2). wPRESS as kNN-average using 5 neighbors (Figure 2) was a less reliable approach for predictive uncertainty assessment (Table 2), and these methods under-estimated the predictive uncertainty in relation to the other methods (Figure 4A and 5).

The double-cross validation succeeded in avoiding over fitting of the PLS model, and generated together with LOO-estimated PRedictive Errors a reasonable Empirical coverage (Figure 4). The algorithm for modeling was judged as trustworthy for the simulation experiments.

The local Empirical coverage for compounds in the test data set (Figure 4) is an estimate of the confidence in each prediction, and a probabilistic measure of reliability (Figure 1).

The probabilistic measure of reliability in prediction is correlated to prediction error (Figure 5).

Figure 3. Predictive errors against similarity measures. The line is the predictive error estimated by PRE

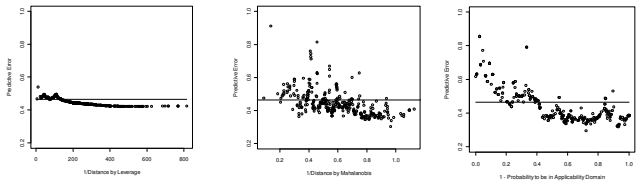


Figure 4. Empirical coverage and Local Empirical coverage

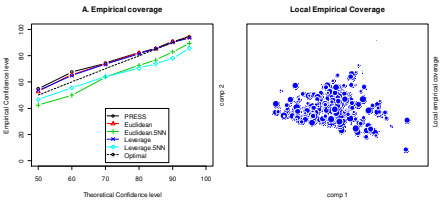
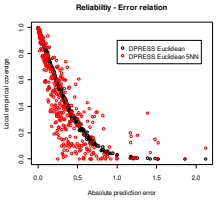


Figure 5.



METHOD

Empirical testing of the performance of using analogy

Based on published QSAR data we did an empirical experiment of the reliability of QSAR predictions derived from different ways to assess predictive error, to answer 1) if wPRESS generates more reliable predictions than PRESS and 2) if there is a similarity measure that performs better than others?

QSAR data

As a case study we evaluated wPRESS for a data set on aqueous toxicity against Tetrahymena pyriformis (pIGC50) previously addressed in another study [3]. Descriptors were generated by Dragon v. 5.4.

A systematic evaluation were done on a suite of data sets containing ten SARs regressions for toxicologically associated assays obtained from PubChem, and 16 congeneric QSAR regression data sets taken from the chemoinformatics web page, as well as data from US-EPA, FDA and NTP. Here we show results for data sets where chemical structure is represented by 177 physio-chemical descriptors of RDkit. However, one of the similarity measures were pairwise Euclidean distances for circular fingerprint as implemented in RDkit with radius 1.

Predictive QSAR models

For each QSAR data set we divided data into training (2/3) and testing (1/3). Optimal PLS model complexity was found by double cross validation. Leave-One-Out cross validated square errors of prediction. For each compound in the test set we assessed the predictive distribution for each type of method to generate wPRESS. All compounds in the test data set as in the AD. This was repeated 10 times with a new partition into training and testing.

Evaluation of PRediction Error Sums

The reliability of PRESS and the wPRESS's were evaluated in the probabilistic framework (Figure 1), by making inference of actual values in the test set under the predictive distribution. The rank of the summed logged likelihood scores for the test set provides a simple way to compare the reliability of alternative methods to assess the predictive distribution. Differences in reliability were tested for using Friedman rank sum test. Posthoc testing aimed to find wPRESS with significantly higher performance in relation to PRESS in assessing predictive error.

Table 3 Congeneric QSAR Regression RDK

	Total Rank	ACE	ACHE	AMPH1	BZR	BZR_S	COX2	DHFR	EDC	ERDK	FDAMDD	HIVPR	HIVRT	HPTP	solubilityTrain	THERM	THR
PRESS	5	4	5	6	4	7	6	4	7	5	4	5	9	7	4	6	3
Euclidean	1	5	1	1	1	2	1	1	3	2	1	1	6	3	1	9	2
Euclidean.5NN	7	6	7	8	8	3	8	7	1	6	7	9	10	10	6	10	8
Mahalanobis	4	1	6	3	5	5	3	6	4	4	5	4	3	5	3	1	6
Mahalanobis.5NN	9	9	9	10	11	10	9	9	10	10	8	7	1	2	9	5	7
FingerPrints	2	3	2	2	2	1	2	2	5	1	2	3	7	4	5	4	1
FingerPrints.5NN	8	10	2	7	5	4	5	8	6	9	10	11	1	9	8	8	11
Leverage	3	2	4	3	3	6	4	3	2	3	3	2	5	6	2	2	4
Leverage.5NN	10	8	11	9	10	8	10	9	9	8	9	8	4	1	10	3	9
Density	6	6	8	5	5	9	7	5	8	7	6	6	8	8	7	7	5
Density.5NN	11	11	10	11	9	11	11	11	11	11	11	10	11	11	11	11	10

Table 4 PubChem Regression RDK

	Small size (100)									Medium size (500)								
	Total Rank	631	639	677	682	832	1239	1479	1511	2156	2796	Total Rank	631	639	677	682	832	1239	1479	1511	2156	2796
PRESS	2	2	3	1	2	5	1	4	2	3		3	4	2	3	3	3	3	3	4	3	
Euclidean	1	1	4	6	1	1	1	5	2	1	4	1	1	5	1	1	1	1	1	1	1	1
Euclidean.5NN	7	8	9	11	2	5	7	11	9	8	6	10	11	11	11	8	11	7	7	7	8	10
Mahalanobis	5	5	5	5	4	7	4	6	1	5	7	5	5	4	5	6	6	5	4	6	5	4
Mahalanobis.5NN	8	9	8	8	9	8	11	9	5	9	9	8	8	8	8	8	11	8	8	8	9	10
FingerPrints	4	3	6	2	5	4	3	3	2	7	4	1	4	2	6	4	4	4	3	10	4	2
FingerPrints.5NN	10	11	11	7	7	11	10	7	11	10	8	7	7	10	7	7	7	10	6	9	7	7
Leverage	3	3	1	4	2	3	2	4	3	3	5	2	3	1	1	2	2	2	2	2	2	2
Leverage.5NN	9	7	8	9	10	9	9	10	5	7	10	9	8	9	9	10	9	8	8	9	9	9
Density	6	6	1	3	7	6	6	3	7	6	1	6	6	3	6	5	5	6	5	5	5	6
Density.5NN	11	10	7	10	11	10	8	8	10	11	11	11	10	7	10	9	10	11	11	7	10	11

Table 5 PubChem Regression RDK

	Small (100)				Medium (500)				Large (1446-2302)			
	631	639	677	1479	631	639	677	1479	631	639	677	1479
PRESS	2	3	1	1	4	2	3	3	5	4	4	5
Euclidean	1	4	6	5	1	5	1	1	1	2	1	1
Euclidean.5NN	8	9	11	11	11	11	11	7	10	11	11	7
Mahalanobis	5	5	5	6	5	4	5	4	8	5	5	2
Mahalanobis.5NN	9	8	8	8	8	8	8	8	6	8	8	7
FingerPrints	3	6	2	2	2	6	4	10	4	1	3	10
FingerPrints.5NN	11	11	7	7	7	10	7	6	9	10	7	4
Leverage	3	1	4	4	3	1	1	2	2	3	2	3
Leverage.5NN	7	8	9	10	8	9	9	8	7	8	9	9
Density	6	1	3	3	6	3	6	5	3	6	6	6
Density.5NN	10	7	10	8	10	7	10	11	11	7	10	11

Statistically significant higher performance compared to PRESS in **bold**

EMPIRICAL RESULT

wPRESS based on Euclidean distances and Leverages performed in general as good as or better than PRESS (Table 2 to 5). Similarity based on Fingerprints had a higher performance compared to PRESS for the Congeneric QSAR regression data sets (Table 3). A reason to low performance of Fingerprints for the PubChem data sets could be that the radius of one was to restrictive to detect differences in predictive uncertainty. When reducing the sizes of the data sets, mimicking a situation of a less well specified predictive model, PRESS and the wPRESS based on AD density performed better than before (Table 4 and 5).

CONCLUSIONS

Does assessing predictive uncertainty by analogy make sense?

>>>>>No!
PRESS behaves in most cases as well as wPRESS and is easier to apply.
The robustness of wPRESS is sensitive to the choice of similarity measure.
The performance of different wPRESS varies between data sets.
wPRESS should not be used to correct for unexplained variation in a poor predictive regression model.

>>>>>Yes!
The higher reliability of wPRESS compared to PRESS indicates the presence of heterogeneous variance in regression errors and/or prevailing heterogeneity in predictive errors (variances).
The good performance of wPRESS based on Leverage is supported by theory.
PRESS is a special case of wPRESS.
wPRESS uses LOO and therefore which similarity measure to use can be selected on the external test set.

The results are encouraging for continued development of wPRESS as the Euclidean and Leverage methods almost consistently rank better than PRESS. We will continue evaluating methods to assess the predictive distribution to open up for the possibility to develop approaches to assess reliability in QSAR regressions in a probabilistic framework.

Acknowledgements
This study was funded by the FP7 project CADASTER (grant agreement 212668).
A CADASTER objective is to solve issues on how to practically integrate QSARs model predictions in risk assessment with the long-term goal to increase the use of non-testing information for regulatory decisions, while meeting the main challenges of quantifying and reducing uncertainty.

References
[1] Sahlin U, Filipsson M, Öberg T. A risk assessment perspective of current practice in characterizing uncertainties in QSAR regression predictions. Mol. Inf. 2011;30:551.
[2] Clark R. DPRESS: Localizing estimates of predictive uncertainty. J Cheminf. 2009;1:11.
[3] Tetko IV et al. Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: Focusing on applicability domain and overfitting by variable selection. J. Chem. Inf. Model. 2008;48:1733.

Linnæus University