Chemometrics and Intelligent Laboratory Systems 152 (2016) 18–33

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Be aware of error measures. Further studies on validation of predictive QSAR models☆

Kunal Roy *, Rudra Narayan Das [1], Pravin Ambure [1], Rahul B. Aher [1]

Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700 032, India

ARTICLE INFO

Article history:
Received 18 November 2015
Received in revised form 8 January 2016
Accepted 10 January 2016
Available online 18 January 2016

Keywords:
QSAR
Validation
Error measures
Dispersion
MAE

ABSTRACT

Validation is the most crucial concept for development and application of quantitative structure–activity relationship (QSAR) models. The validation process confirms the reliability of the developed QSAR models along with the acceptability of each step during model development such as assessing the quality of input data, dataset diversity, predictability on an external set, domain of applicability and mechanistic interpretability. External validation or validation using an independent test set is usually considered as the gold standard in evaluating the quality of predictions from a QSAR model. The external predictivity of QSAR models is commonly described by employing various validation metrics, which can be broadly categorized into two major classes, viz., R^2 based metrics namely R^2_{test}, $Q^2_{(\mathrm{ext\_F1})}$, and $Q^2_{(\mathrm{ext\_F2})}$, and purely error based measures like predicted residual sum of squares (PRESS), root mean square error (RMSE), and mean absolute error (MAE). The problem associated with the error based measures is the absence of any well-defined threshold for determining the quality of predictions making the R^2 based metrics more suitable for use due to easy comprehension. However, in this paper, we show the problems associated with the R^2 based validation metrics commonly used in QSAR studies, since their values are highly dependent on the range of the response values of the test set compounds and their distribution pattern around the training/test set mean. We also propose a guideline for determining the quality of predictions based on MAE and its standard deviation computed from the test set predictions after omitting 5% high residual data points in order to obviate the influence of any rarely occurring high prediction errors that may significantly obscure the quality of predictions for the whole test set. In this manner, we try to evaluate the prediction performance of a model on most (95%) of the data points present in the external set. An online tool (XternalValidationPlus) for computing the suggested MAE based criteria (along with other conventional metrics) for external validation has been made available at http://dtclab.webs.com/software-tools and http://teqip.jdvu.ac.in/QSAR_Tools/. The MAE based criteria suggested here along with other commonly used validation metrics may be applied to evaluate predictive performance of QSAR models with a greater degree of confidence.

1. Introduction

Quantitative structure–activity relationship (QSAR) modeling is utilized in rational drug design, environmental risk assessment and fate modeling, toxicity and property prediction of chemicals and pharmaceuticals. A QSAR model represents a mathematical relationship for a set of molecules (*training set*) with a known response (*activity*/*toxicity*/*property*) obtained from application of various chemometric techniques (statistical tools). The actual relationship is built between the structural features of a molecule expressed in quantitative terms (*descriptors*/*independent variables*) that are derived computationally (or experimentally in some cases) and the dependent variable or the response, which should always be experimentally derived. QSAR is a time- as well as cost-effective technique that supports the 3Rs (*replacement, refinement and reduction in animals in research*) paradigm [1]. The chemical and toxicological regulatory agencies worldwide have been employing QSAR models for decision-making frameworks in risk and safety assessments for a number of years [2].

Validation is the most crucial concept for development and application of any QSAR model. The validation process confirms the reliability of the developed QSAR model along with the acceptability of each step during model development such as assessing the quality of input data, dataset diversity, predictability on external set, domain of applicability and mechanistic interpretability. For regulatory acceptance of QSAR models, five guidelines are agreed by the Organization for Economic Co-operation and Development (OECD) [3], and these cover the following criteria: (i) a defined endpoint, (ii) an unambiguous algorithm, (iii) the domain of applicability, (iv) appropriate measures of goodness of fit, robustness and predictivity of the developed model, and (v) a

☆ Presented in part at the 23rd Conference on Current Trends in Computational Chemistry (23rd CCTCC) held in Jackson, Mississippi, USA, on November 13–14, 2015.

* Corresponding author. Tel.: +91 98315 94140; fax: +91 33 2837 1078.
E-mail addresses: kunalroy_in@yahoo.com, kroy@pharma.jdvu.ac.in (K. Roy).
URL: http://sites.google.com/site/kunalroyindia/ (K. Roy).

[1] Tel.: +91 98,315 94,140; fax: +91 33 2837 1078.

http://dx.doi.org/10.1016/j.chemolab.2016.01.008
0169-7439/© 2016 Elsevier B.V. All rights reserved.

mechanistic interpretation, if possible. These guidelines are now referred to as OECD Principles for the validation of QSARs. As a result, different groups of researchers have shown their keen interest towards the development of more appropriate validation metrics for precise and predictive QSAR model development [3–8].

Recently, Alexander et al. [9] have suggested some shortcomings in the model fit criteria previously suggested by Golbraikh and Tropsha, [5] and further proposed that only two simple criteria might be sufficient for judging model usefulness: high R^2 (*correlation coefficient*) and low root mean square error (RMSE) of test set predictions. In the present paper, we have tried to relook the problem in a greater detail. We have highlighted some problems associated with the conventional R^2 based validation metrics and suggested a set of criteria based on model errors for unbiased judgment of the quality of model predictions. Although the problems associated with the R^2 based formalism were identified long back [10], QSAR practitioners usually rely on these metrics for evaluating the predictive potential of models. Therefore, we aim to provide here the readers with an overview of the shortcomings of the conventional R^2 based metrics as applied in QSAR studies and also encourage easy interpretation of the quality of predictions from the error based judgment.

2. Problems with the conventional metrics

Validation of QSAR models is a crucial issue for judging their ability of prediction for the chemicals not employed during model development. In consonance with the OECD guidelines regarding the model fitness, robustness as well as predictivity, a number of statistical metrics are used by different research groups in this field. In this article, we shall restrict our discussion to the validation aspect involving test set compounds only, i.e., metrics characterizing external validation of a model.

A commonly used regression based measure is determination coefficient (R^2) between the observed and predicted response values of the test set compounds. This metric may be computed based on the following expression [11].

$$R^2 = \frac{\left[\sum\left\{\left(Y_{obs}-\overline{Y_{obs}}\right)\times\left(Y_{pred}-\overline{Y_{pred}}\right)\right\}\right]^2}{\sum\left(Y_{obs}-\overline{Y_{obs}}\right)^2\times\sum\left(Y_{pred}-\overline{Y_{pred}}\right)^2} \tag{1}$$

In Eq. (1), Y_{obs} and Y_{pred} correspond to the observed (i.e., experimental) and predicted response values respectively of the test set compounds. Instead of providing a true picture of the prediction errors encountered, the R^2 metric as defined in Eq. 1 attempts to provide a relative pattern of changes in the values of the observed response with respect to the predicted ones. As a result, this metric can furnish acceptable values for a constant magnitude of errors for all the samples even if it is very high. A way out to overcome this problem may be the use of the regression through origin (RTO) approach where the best fitted line is deliberately forced through origin ($Y_{obs} = 0$, $Y_{pred} = 0$) in order to penalize the R^2 value obtained from the corresponding normal regression analysis in case of large prediction errors [11]. Based on the judgment of RTO derived method, researchers in this field have formulated model validation criteria such as Golbraikh and Tropsha's criteria [6] as well as different r_m^2 metrics [7,12,13].

However, the RTO approach is able to identify prediction errors of a model as long as the data are devoid of any 'systematic error' and/or model bias. Systematic error is usually characterized by bias in model predictions. However, such errors in analytical experiments are avoidable and mostly represent those arising from operational perspective of the analyst, instrumental adjustments, reagent based defects, as well as improper method based flaws giving inaccurate results [14]. Dearden et al. [15] have identified such errors in models due to improper selection of model variables. A biased model prediction may be characterized by all error values of same sign, i.e., all (or disproportionately high fractions) being positives or all (or disproportionately high fractions) being negatives. The determination coefficient R^2 and its origin based counterpart, i.e., R_0^2 are applicable only if the predicted data are devoid of such existing 'systematic error' feature, otherwise it might give a wrong assessment of the model predictivity. In case of the presence of any systematic error or model bias for a particular test set, attempt should be made to change the model to remove the systematic error as such test set is not suitable for predictions from the developed model in any validation experiment. This is something similar to adjusting instrumental error before doing an instrumental analysis and such error has nothing to do with the quality of determinations (predictions in our case). Some methods for the identification of systematic error include residual plot analysis [15], implementation of Kriging models [16], comparison analysis involving average error and average absolute error measure [17].

The external predictivity of QSAR models is commonly described by employing various validation metrics, and these can be broadly categorized into two major classes, viz., R^2 based metrics namely $Q^2_{ext(F1)}$ and $Q^2_{ext(F2)}$ [18] and error based measures like predicted residual sum of squares (PRESS), standard error of estimate (SEE), root mean square error (RMSE), and mean absolute error (MAE) [19]. An alternative general formula for R^2 has been furnished in Eq. 2 which is most commonly used for computation of different R^2 based validation metrics (or Q^2_{ext}).

$$R^2 = 1-\frac{\sum\left(Y_{obs}-Y_{pred}\right)^2}{\sum\left(Y_{obs}-\overline{Y}\right)^2} \tag{2}$$

In Eq. 2, the experimental and predicted response values of a chemical have been designated using Y_{obs} and Y_{pred} respectively, while $\overline{Y}$ represents the mean response value of the training set or the test set compounds, depending upon the metric used (for example, $Q^2_{ext(F1)}$ and $Q^2_{ext(F2)}$). $Q^2_{ext(F1)}$ uses the training set mean value while it is the test set mean in the case of $Q^2_{ext(F2)}$.The numerator of the fraction shown in Eq. 2 is a measure of prediction error and the R^2 metric measures the model performance (in terms of prediction errors) in comparison to the performance of a "no model" situation (that is the mean of the response values of the training or test set compounds considered as the reference). A model will be of no use if its prediction performance is not better than, at least, the performance of the mean (i.e., "no model"). Thus, a model may be considered acceptable when the values of these R^2 based metrics ($Q^2_{ext(F1)}$ and $Q^2_{ext(F2)}$) are at least more than 0.5; the closer are the values to unity, the greater is the confidence in prediction precisions. Although this formalism seems logical, the results from the comparison with the performance of the mean can sometimes be misleading since it is greatly influenced by the range of the corresponding training/test set data, the average value of which is used in Eq. 2. Another important aspect for the over- and under-estimation of prediction errors by the Q^2_{ext} metrics is the distribution of the response data around mean. In the case of the test set mean based assessment, i.e., $Q^2_{ext(F2),}$ if most of the response data points remain in close neighborhood of the mean value leaving only a low fraction away from it, the mean can perform well as an estimate of the individual responses and the value of the metric can be low in spite of the presence of low amount of prediction errors. Thus, the judgment provided by the Q^2_{ext} metrics is not only dependent on model based predictions but also on other factors like range as well as distribution of the response data around mean, and therefore such metrics cannot be identified as sufficient measures for external validation. We may mention here that the expression of R^2 for external validation as suggested by Alexander et al. [9] actually corresponds to $Q^2_{ext(F2)}$, though they did not mention about it explicitly in their paper.

While the Q^2_{ext} metrics may give misleading results regarding the quality of predictions, prediction error based metrics like PRESS, SEE, MAE, and RMSE [19] give more straight-forward results. Now, the main problem while dealing with such metrics is the absence of a suitable threshold value unlike the Q^2_{ext} metrics. It is to be noted that