Journal of
Cheminformatics

ORAL PRESENTATION Open Access

Quantifying model errors using similarity to training data

Rob D Brown[1*], JD Honeycutt[1], SL Aaron[2]

From 5th German Conference on Cheminformatics: 23. CIC-Workshop
Goslar, Germany. 8-10 November 2009

When making a prediction with a statistical model, it is not sufficient to know that the model is "good", in the sense that it is able to make accurate predictions on test data. Another relevant question is: How good is the model for a specific sample whose properties we wish to predict? Stated another way: Is the sample within or outside the model's domain of applicability or what is the degree to which a test compound is within the model's domain of applicability. Numerous studies have been done on determining appropriate measures to address this question [1-4]. Here we focus on a derivative question: Can we determine an applicability domain measure suitable for deriving quantitative error bars – that is, error bars which accurately reflect the expected error when making predictions for specified values of the domain measure? Such a measure could then be used to provide an indication of the confidence in a given prediction (i.e. the likely error in a prediction based on to what degree the test compound is part of the model's domain of applicability).Ideally, we wish such a measure to be simple to calculate and to understand, to apply to models of all types – including classification and regression models for both molecular and non-molecular data - and to be free of adjustable parameters. Consistent with recent work by others [5,6], the measures we have seen that best meet these criteria are distances to individual samples in the training data. We describe our attempts to construct a recipe for deriving quantitative error bars from these distances.

Author details
[1]Accelrys Inc, 10188 Telesis Court, San Diego, CA 92121, USA. [2]Accelrys Inc, Cambridge, UK.

Published: 4 May 2010

References

1. Eriksson L, Jaworska J, Worth AP, Cronin MTD, McDowell RM, Gramatica P: **Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs.** *Environmental Health Perspectives* 2003, **111**:1361.
2. Tropsha A, Gramatica P, Gombar VK: **The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models.** *QSAR Comb Sci* 2003, **22**:69.
3. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T: **QSAR applicabilty domain estimation by projection of the training set descriptor space: a review.** *Altern Lab Anim* 2005, **33**:445-59.
4. Stanforth RW, Kolossov E, Mirkin B: **A Measure of Domain of Applicability for QSAR Modelling Based on Intelligent K-Means Clustering.** *QSAR & Combinatorial Science* 2007, **26**:837-.
5. Sheridan RP, Feuston BP, Maiorov VN, Kearsley SK: **Similarity to Molecules in the Training Set Is a Good Discriminator for Prediction Accuracy in QSAR.** *J Chem Inf Comput Sci* 2004, **44**:1912.
6. Horvath D, Marcou G, Varnek A: **Predicting the Predictability: A Unified Approach to the Applicability Domain Problem of QSAR Models.** *J Chem Inf Comput Sci* 2009, **49**:49.

doi:10.1186/1758-2946-2-S1-O7
Cite this article as: Brown *et al.*: **Quantifying model errors using similarity to training data.** *Journal of Cheminformatics* 2010 **2**(Suppl 1):O7.

[1]Accelrys Inc, 10188 Telesis Court, San Diego, CA 92121, USA