Sushko *et al. Journal of Cheminformatics* 2010, **2**(Suppl 1):P41
http://www.jcheminf.com/content/2/S1/P41

Journal of Cheminformatics

POSTER PRESENTATION Open Access

Applicability domain for classification problems

Iurii Sushko*, S Novotarskyi, AK Pandey, R Körner, Igor Tetko

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

Classification models are frequent in QSAR modeling. It is of crucial importance to provide good accuracy estimation for classification. Applicability domain provides additional information to identify which compounds are classified with best accuracy and which are expected to have poor and unreliable predictions. The selection of the most reliable predictions can dramatically improve performance of methods while decreasing coverage of predictions [1].

In binary classification problems, labels for machine learning methods are discrete {-1, 1}. Nonetheless, model usually yields prediction that is continuous. Most apparent metrics for accuracy estimation is distance between prediction point and edge of a class, i.e. the more is the distance between prediction the edge of the class, the more reliable and accurate is the prediction of given compound. This metric has been already used in several previous studies (e.g., [2]) and demonstrated good separation of reliable and non-reliable classifications. In quantitative predictions, the standard deviation of ensemble predictions has been found as the most accurate measure distance in a recent benchmarking [3].

We propose to integrate both metrics. Rather than giving a point estimate, this approach provides us with a probability distribution of finding particular compound in one of the classes. Suggested metrics is probability

$$\int_E N(a,v,x)dx$$

where E is class domain a - ensemble's average prediction, v – variance of ensemble's prediction, $N(a, v, x)$ is probability density of the Gaussian distribution. Performance of this metric and its comparison to the traditional ones are evaluated for several QSAR/QSPR classification problems. The developed approach can be freely accessed to develop and estimate applicability domain of classification models at http://qspr.eu web site.

Sushko, I., Helmholtz Zentrum München/IBIS, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany

Published: 4 May 2010

References

1. Tetko IV, Bruneau P, Mewes HW, Rohrer DC, Poda GI: *Drug Discov Today* 2006, **11**:700.
2. Manallack DT, Tehan BG, Gancia E, Hudson BD, Ford MG, Livingstone DJ, Whitley DC, Pitt WR: *J Chem Inf Comput Sci* 2003, **43**:674.
3. Tetko IV, Sushko I, Pandey AK, Zhu H, Tropsha A, Papa E, Oberg T, Todeschini R, Fourches D, Varnek A: *J Chem Inf Model* 2008, **48**:1733.

doi:10.1186/1758-2946-2-S1-P41
Cite this article as: Sushko *et al.*: Applicability domain for classification problems. *Journal of Cheminformatics* 2010 **2**(Suppl 1):P41.