[P23] Comparison of Different Measures for the Domain of Applicability of Classification Models

Miriam Mathea, Waldemar Klingspohn, Knut Baumann

Institute of Medicinal and Pharmaceutical Chemistry, Braunschweig University of Technology, Beethovenstr. 55, 38106 Braunschweig, Germany

Quantitative Structure-Activity Relationship (QSAR) models are widely used to predict the biological activity and various other properties of drug candidates. The performance of the employed model is routinely characterized by the prediction error (PE) which is estimated with test set molecules not involved in model building and model selection (so-called external test set). While test set prediction provides a global measure of the prediction error, the individual prediction error for a single test set molecule is far more interesting for decision making in the drug development process.

To avoid gross errors, for validated models the so-called domain of applicability should be rigorously defined. The domain of applicability is defined as the “response and chemical structure space in which the model is considered to make predictions with a given reliability, in order to express the scope and limitations of a model” [1]. A perfect measure to characterize the domain of applicability would separate the response and chemical space into islands with a defined upper bound on the prediction error and the remaining space where the prediction error exceeds this upper bound. Hence, such a perfect measure would allow to identify molecules for which predictions are reliable (on the islands) or unreliable (remainder). Unfortunately, such a perfect measure is rarely available. Even more practically relevant but even harder to define would be a measure that correlates well with the size of the prediction error (i.e. small size of the measure corresponds to a small PE and vice versa). These measures are often called distance to model measures [2]. Once again, a perfect distance to model measure is rarely available. Yet, even measures that indicate a trend are of huge practical importance.

The aim of this study is to systematically evaluate different measures for the domain of applicability for classification models to identify those that correlate best with the PE of an individual molecule where the PE is expressed as probability of misclassification of a particular molecule. In a full factorial design the four classification techniques support vector machines, linear discriminant analysis, k-nearest neighbor classification, and random forests are evaluated in combination with various distance to model measures in order to rank these measures for every method and to identify matching pairs that perform best for a large mutagenicity benchmark data set [3].

[1] T.I. Netzeva et al. Altern Lab Anim 33 (2005) 155-173
[2] I. Sushko et al. J. Chem. Inf. Model. 50 (2010) 2094-2111
[3] Hansen et al. J. Chem. Inf. Model. 49 (9) (2009) 2077-81