Regulatory Toxicology and Pharmacology 71 (2015) 279–284

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

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

Introducing conformal prediction in predictive modeling for regulatory purposes. A transparent and flexible alternative to applicability domain determination

Ulf Norinder [a,*], Lars Carlsson [b], Scott Boyer [a], Martin Eklund [b,c]

[a] *Swedish Toxicology Sciences Research Center, SE-151 36 Södertälje, Sweden*
[b] *AstraZeneca Research and Development, SE-431 83 Mölndal, Sweden*
[c] *Dept. Surgery, University of California at San Francisco (UCSF), San Francisco, CA 94115, USA*

ARTICLE INFO

Article history:
Received 29 October 2014
Available online 2 January 2015

Keywords:
Applicability domain
CAESAR
Confidence predictor
Conformal prediction
MOE descriptors
Conformity score
Random forest
REACH
Signature descriptors

ABSTRACT

Conformal prediction is presented as a framework which fulfills the OECD principles on (Q)SAR. It offers an intuitive extension to the application of machine-learning methods to structure–activity data where focus is on predictions with pre-defined confidence levels. A conformal predictor will make correct predictions on new compounds corresponding to a user defined confidence level. The confidence level can be altered depending on the situation the predictor is being used in, which allows for flexibility and adaption to risks that the user is willing to take. We demonstrate the usefulness of conformal prediction by applying it to 2 publicly available CAESAR binary classification datasets.

1. Introduction

REACH is the European legislation for safe use of chemicals, which requires information for all chemicals that are currently on the market in Europe in quantities above one tonne per year (REGULATION (EC) No 1907/2006). This is a large initiative and a huge amount of data on each compound is required. To facilitate such data collection alternative methods to direct experimentation, e.g. *in silico* models, have been identified as possible sources for this information. REACH mandates that for *in silico* models, and in particular QSAR models, these models' prediction performance should be validated according to a set of procedures including both internal and external validation. OECD has, in a similar manner, identified a transparent validation process and objective determination for the *reliability* of (Q)SAR models to enhance the regulatory acceptance of such models (OCED, 2004). In November 2004, an agreement was reached among the OECD member countries regarding the principles for validating (Q)SAR models for their regulatory use in the assessment of chemical safety. The agreed principles provide a basis for evaluating regulatory applicability of (Q)SAR models and are published in the OECD document "OECD principles for the Validation, for Regulatory Purpose, of (Q)SAR Models" (Joint Meeting of the Chemicals Committee and Working Party on Chemicals, Pesticides and Biotechnology, 2004). This work was followed up in February 2007 with a document from the Expert Group on (Q)SARs when the OECD published a "Guidance Document on the Validation of (Q)SAR Models" to provide guidance on how specific (Q)SAR models can be evaluated with respect to the OECD principles including a check list for the validation (Joint Meeting of the Chemicals Committee and Working Party on Chemicals, Pesticides and Biotechnology, 2007). Principle 3, among five principles identified as important for the consideration of a (Q)SAR model for regulatory purposes, specifies "a defined domain of applicability". It is further stated, in ANNEX A, of that document that "The need to define an applicability domain (Principle 3) expresses the fact that (Q)SARs are reductionist models which are inevitably associated with limitations in terms of the types of chemical structures, physicochemical properties and mechanisms of action for which the models can generate reliable predictions." and that "Further work is recommended to define what types of information are needed to define (Q)SAR applicability domains, and to develop appropriate methods for obtaining this information" which highlights two crucial requirements for

[*] Corresponding author.
E-mail address: ulf.norinder@swetox.se (U. Norinder).

http://dx.doi.org/10.1016/j.yrtph.2014.12.021
0273-2300/© 2015 Elsevier Inc. All rights reserved.

obtaining reliable predictions, namely, the type of information needed and the development of an appropriate method for obtaining that information.

Several attempts have been made, mainly divided into two categories – one more structure (similarity) focused and the other more information oriented, to fulfill the OECD principles on (Q)SAR predictions, but no one has been able to present well defined estimates on reliability and some have even suggested that there is a discrete cutoff, based on similarity or similar metrics, to when a certain model provides reliable predictions or not (Eriksson et al., 2003; Dimitrov et al., 2005; Netzeva et al., 2005; Bassan and Worth, 2007; Schroeter et al., 2007; Weaver and Gleeson, 2008; Dragos et al., 2009; Sushko et al., 2010; Sahigara et al., 2012; Sheridan, 2012, 2013; Keefer et al., 2013; Wood et al., 2013). All these attempts to AD estimations are built on the assumed ability of an AD metric to capture ‘nearness’: relative proximity (according to the AD measure) equates to assumed relatively higher accuracy. This makes intuitive sense, but crucial questions (and the genesis of the ambiguity) for the application of the AD concept are ‘How close is close enough for an accurate enough prediction? And according to what AD metrics?’

What we ideally would like to know is in fact that a particular prediction is within a certain interval (prediction region) with a given (user specified) confidence (compare with the notion of a confidence interval). In the following section we describe conformal prediction as a method for achieving this and as a solution to the problem of ambiguity related to applicability domain estimation. We apply the method to 2 publicly available CAESAR binary classification datasets as an illustration of the method.

2. Materials and methods

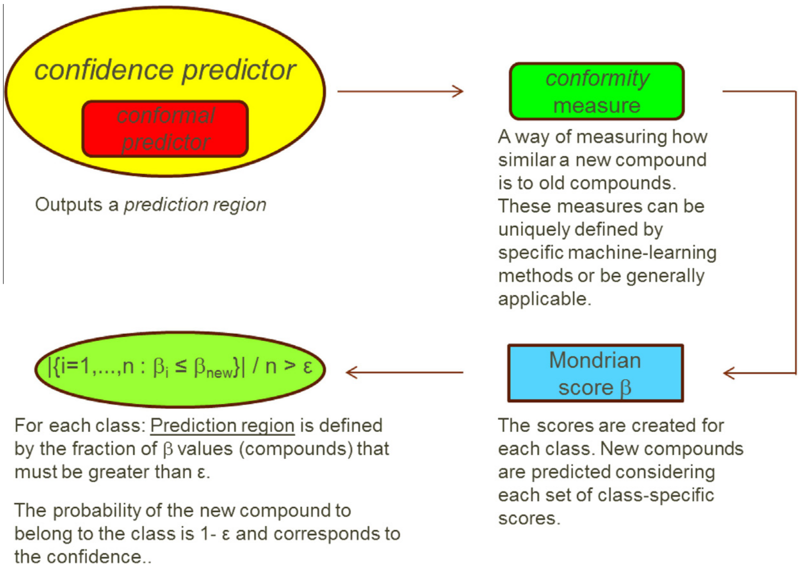
2.1. Conformal prediction

We will in this section introduce the conformal predictor, with a focus on informally explaining the idea behind it. We will do this graphically (Fig. 1) and by using an example and pseudo code to illustrate how to use conformal predictions. For a more formal description and for proofs of the mathematical theorems on which the conformal prediction framework is built, we refer to Vovk et al. (2005). For some initial, more mathematically oriented, work in the QSAR domain, we refer to Eklund et al. (2012, 2013). We also refer to Norinder et al. (2014) for a more chemoinformatics oriented description.

A *confidence predictor* is a prediction algorithm that outputs a prediction region, which contrasts to the single label (classification) predictions output by standard prediction algorithms, e.g. support vector machines (SVM) or random forest (RF). A *conformal predictor* is a particular type of confidence predictor. A confidence predictor is said to be *valid* if the frequency of errors it commits does not exceed ε at a chosen confidence level $1 - \varepsilon$, and *efficient* if the prediction regions it outputs are as small as possible, i.e. predicting mostly single classes in classification problems. Conformal predictors have the attractive property of always being valid, under the assumption that compounds are independently drawn from the same distribution (this assumption is also made for most standard prediction algorithms used in QSAR, e.g. SVM or random forest, so conformal prediction does not introduce new assumptions in addition to the ones we generally use already for QSAR modeling).

To construct a conformal predictor’s prediction regions, we need to define a *conformity measure*. Intuitively, this is a way of measuring how similar a new compound is to existing (old) compounds (a conformity score thus serves the same purpose as an AD measure, and most AD measures can be used as conformity scores). Relating conformity scores of compounds to be predicted with conformity scores of previously experimentally tested compounds is the core of conformal predictors. We do this using a *p*-value, the number of existing compounds that have as small or smaller conformity scores as the new compound, divided by the total number of compounds. If this value (fraction) is small compared to the values for existing compounds then the new compound is very non-conforming, i.e. the new compound is different from previous compounds because of its different conformity score compared to most of the existing compounds in the training set. On the other hand, if the value (fraction) is large compared to the values for existing compounds, then the new compound is very conforming, i.e. very similar to most of the existing compounds in the training set.

To evaluate the predicted class(es) of new compounds from the model a significance level (ε) is set, e.g. at 0.2 (corresponding to a 0.8 confidence level), that is appropriate for the modeling situation. For each new compound the fraction of conformity scores for existing compounds less than the conformity score of the new


confidence predictor
conformal predictor
Outputs a *prediction region*
conformity measure
Mondrian score β
$|\{i=1,\ldots,n : \beta_i \le \beta_{new}\}| / n > \varepsilon$

A way of measuring how similar a new compound is to old compounds. These measures can be uniquely defined by specific machine-learning methods or be generally applicable.

The scores are created for each class. New compounds are predicted considering each set of class-specific scores.

For each class: Prediction region is defined by the fraction of β values (compounds) that must be greater than ε.

The probability of the new compound to belong to the class is 1- ε and corresponds to the confidence..

Fig. 1. Conformal predictors. Flow chart showing the procedure for obtaining the class CP probabilities of new compounds.