QUANTITATIVE ESTIMATION OF PREDICTIVE UNCERTAINTY FOR ENSEMBLE MODELS

Robert D. Clark, Marvin Waldman, and Robert Fraczkiewicz

Simulations Plus, Inc., 42505 10th Street West, Lancaster CA 93534

Introduction

The performance of QSAR models has traditionally been evaluated in terms of aggregate statistics – sensitivity, specificity, root mean square error (RMSE), R^2, etc. – for some kind of test set. More recently, the fact that models are generally more reliable for compounds that are similar to those included in the training set than for those which are dissimilar has led to the concept of "applicability domain." A simple binary categorization of a compound as being inside or outside of a model's applicability domain is often too coarse for regulatory purposes, however, and often for lead optimization purposes as well. One way to address this shortcoming is to find ways to relate the degree of consensus among the multiple predictions within an ensemble model to the degree of error expected for the consensus prediction, i.e., to estimate the predictive uncertainty. We have found that overdisperse distributions can be used to do this: beta binomials for ensemble classification models and gamma distributions for ensemble regression models.

Methods

Artificial neural network ensemble (ANNE) classification and regression models were built in ADMET Modeler™ using 2D molecular property descriptors. All networks within an ensemble (ANNE) model share a common set of input descriptors and have the same number of neurons, but each ANN is trained on a separate partition of the training pool into training and verification sets (an external test set was held out of training). A network's performance on its verification set was monitored to prevent overtraining.

Predictions for ANNE classification models were obtained by counting the number of positive network "votes" and comparing that vote tally to a threshold. A beta binomial F was fitted to the *cumulative distribution of errors* across the number (tally) k of networks in the ensemble that cast a positive "vote" for that prediction. A second beta binomial G was fitted to the *cumulative distribution of predictions* across the tally of positive votes.

For a prediction based on k positive votes, the uncertainty $u(k)$ is then given by:

$$u(k) = ER \times f(k)/g(k) \qquad \text{(Eqn 1)}$$

An example analysis is shown in Figure 1. See Clark *et al.** for details.

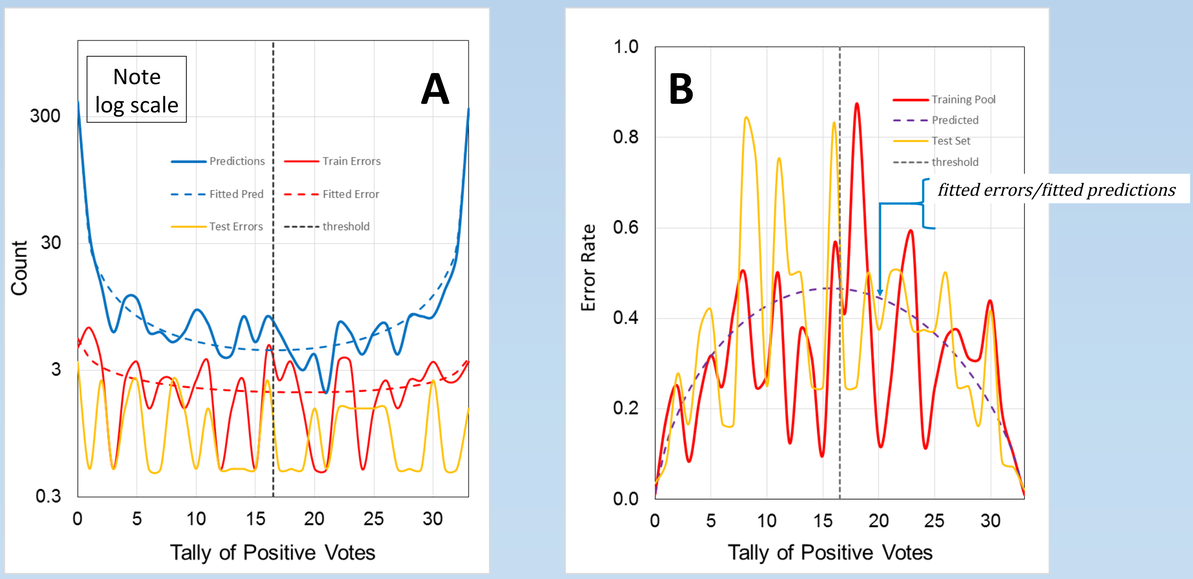

Figure 1. Beta binomial uncertainty analysis for a logP classification model (≤ 2 vs > 2) based on a training pool of 969 compounds. The ensemble was composed of 33 networks, each of which had 40 inputs and 6 neurons. (**A**) The solid blue line shows the distribution of training pool predictions, the solid red line shows the distribution of training pool errors, and the orange line shows the distribution of errors for an external 242 compound test set. The red and blue dashed lines represent the corresponding fitted beta binomials f and g. (**B**) The solid red and orange lines show the error rate as a function of positive vote tally for the training pool and test set, respectively. The dashed black line shows the uncertainty profile calculated from the fitted beta binomials from **A** using Equation 1.

Figure 2. Gamma uncertainty analysis for an ANNE regression model of aqueous solubility (logS, where S is expressed in M units) based on a training pool of 743 compounds. The ensemble was composed of 33 networks, each of which had 37 inputs and 3 neurons. (**A**) The red points show the cumulative distribution of squared training pool errors and the blue dashed line represents the fitted gamma function F. (**B**) The black dots show the cumulative distribution of squared training pool errors and the green dashed line represents the fitted gamma distribution G.

* Clark *et al.* Using beta binomials to estimate classification uncertainty for ensemble models. *J Cheminfo* 2014, **6**, 34.

Predictions y for ANNE regression models were obtained by averaging the predictions of their constituent networks and the standard deviation s of each prediction was used as a measure of network consensus across the ensemble. A small offset s_0 was introduced because there is a practical limit on how close s can get to 0 but no theoretical one.

A gamma distribution F was fitted to the cumulative distribution of squared errors (ε^2) across the adjusted standard deviation (s-s_0) of the ensemble predictions. A second gamma distribution G was fitted to the cumulative distribution of s itself (e.g., Figure 2). The estimated standard error σ_i for prediction y_i was then calculated from the corresponding regression probability density functions f and g and the overall root mean square error (RMSE):

$$\sigma_i = RMSE \times (f(s_i\text{-}s_0)/g(s_i\text{-}s_0))^{1/2} \qquad \text{(Eqn 2)}$$

Note: F and G share a common scaling factor γ.

Validations

For both examples cited above, the training pool was drawn from a much larger data set – 12580 for the logP classification model and 3595 for the logS regression model. Figures 3 below shows the result of applying Equation 1 to the 11370 held-out logP data points and Figure 4 shows the result of applying Equation 2 to held-out logS data points.

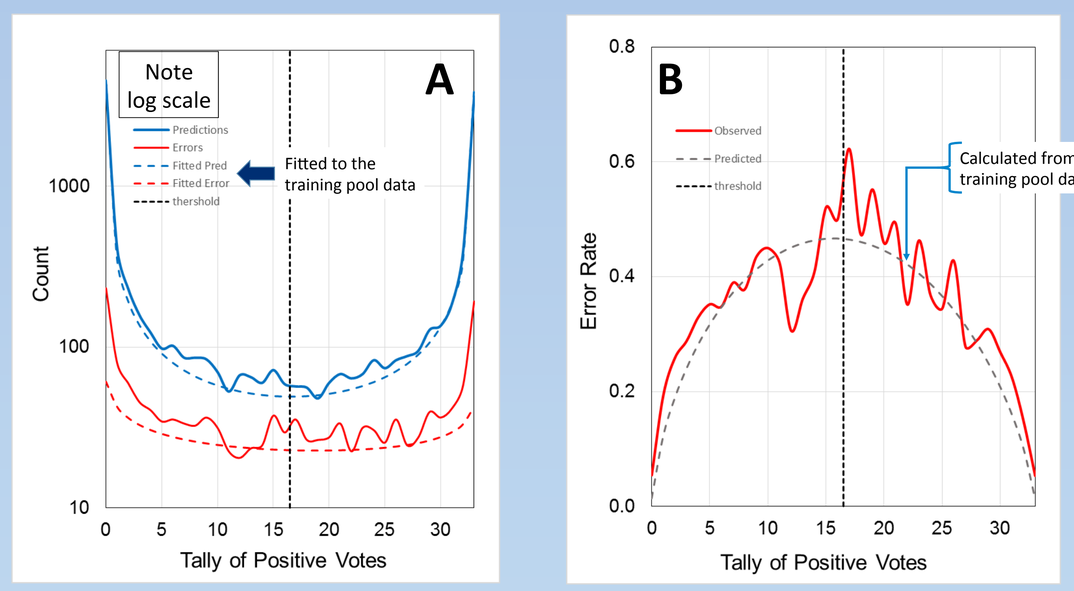

Figure 3. Validation of uncertainty (error rate) estimates for the logP classification model using the 11370 data points held out from the classification model building process. (**A**) The solid red and blue lines show the distribution of validation set errors and predictions, respectively. The dashed lines represent the corresponding fitted beta binomials f and g from the training pool (Figure 1). (**B**) The solid red line shows the error rate as a function of positive vote tally for the validation set, whereas the dashed gray line shows the uncertainty profile calculated from the training pool data (Equation 1 and Figure 1).

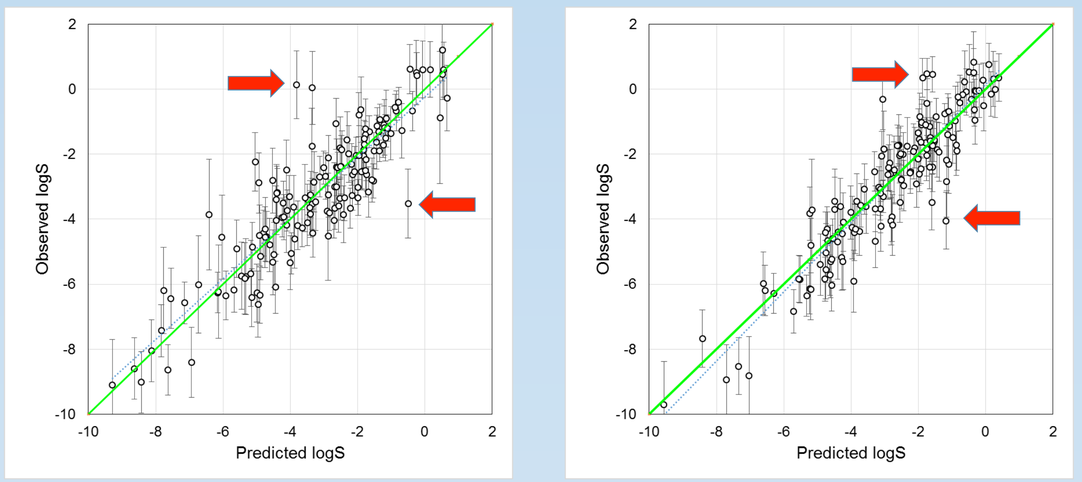

Figure 4. Uncertainty (standard error) estimates for the logS regression model. Results are shown for two 150-compound samples drawn at random from the 2722 compounds in the held-out validation set. Error bars are set equal to ±σi calculated from equation 2. The green line represents a perfect reproduction of the observed data. Red arrows indicate serious outlier points.

The most serious outliers – those where the residual error is larger than would be expected by chance – were examined in detail for the entire data set. Most of these turn out to be cases where the input data is inappropriate. Some are liquids (miscibility is not the same as solubility) and some are surfactants (for which solubility is generally not well-defined). Perhaps the most interesting "bad" predictions are those for compounds where the nominal "solubility" is so high (e.g., 711 g/L for methane disulfonic acid) that classifying water as the solvent becomes risky. Adding a hydrophilic group to such a compound increases the solubility of water *in it*, so the "solubility" of the analog goes *down*!

Conclusions

- Beta binomial and gamma uncertainty analysis make it possible to use the degree of consensus among classification and regression predictions to accurately estimate the reliability of classification and regression predictions from ensemble models, making it possible to distinguish the good and the bad from the downright ugly.
- More precise estimates are easy to identify and the most dubious ones can be discarded in a consistent and well-defined fashion.
- Precise but very inaccurate predictions in the training set – those where most or all of the networks in the ensemble are wrong - are highlighted for further investigation.