Estimating Classification Uncertainty for Ensemble Models

Robert D. Clark, Wenkel Liang, Robert Fraczkiewicz and Marvin Waldman
Simulations Plus, Inc.
Lancaster, CA 93536 USA
Outline

• Motivation
• Model building in ADMET Modeler™
• Binomial, beta and beta-binomial distributions for modeling error distributions
• Fitting an uncertainty profile to the training pool
• Applying an uncertainty profile to the test set
• Using averaging instead of voting
Outline

• Motivation
• Model building in ADMET Modeler™
• Binomial, beta and beta-binomial distributions for modeling error distributions
• Fitting an uncertainty profile to the training pool
• Applying an uncertainty profile to the test set
• Using averaging instead of voting
Motivation

Drug discovery and development are a lot like poker. You cannot win consistently by being lucky. You can win consistently by knowing your opponent (Mother Nature) and by knowing the prospective odds for any given hand.

“You’ve got to know when to hold ‘em, know when to fold ‘em, Know when to walk away, and know when to run. You never count your money when you’re sitting at the table There’ll be time enough for counting when the dealing’s done."

- from “The Gambler” by Kenny Rogers
More Motivation

• Drug discovery & development costs continue to rise
• Quantitative structure-activity relationships (QSARs) have the potential to speed development and reduce costs
• Regulatory agencies support the use of QSARs to guide some decisions
• Considerable progress has been made on how to accurately estimate prospective QSAR predictivity

BUT

• QSAR work has, until recently, focused on assessing the aggregate reliability of QSAR prediction rather than on the reliability of prospective predictions for individual compounds
• Researchers and regulators need to make decisions about individual compounds
Confidence Estimates in ADMET Predictor™ 6.5

experimental Ames classification
Some Relevant Previous Work on Ensemble Predictivity

  - used the variance in artificial neural net ensembles to estimate uncertainty
  - review of uncertainty estimation methods for QSAR
- S. Weaver & M.P. Gleeson. *J Molec Graph Model* **2008**, 26, 1315–1326
  - estimated accuracies of individual regression predictions
  - uncertainty and risk assessment
  - consensus models for *in silico* Ames testing
  - using variance across random forest predictions to help assess confidence
  - confidence metric based on nearest neighbor consensus
Outline

• Motivation

• Model building in ADMET Modeler™
  • Binomial, beta and beta-binomial distributions for modeling error distributions
  • Fitting an uncertainty profile to the training pool
  • Applying an uncertainty profile to the test set
  • Using averaging instead of voting
How We Build Our Ensemble Models

1. Structures and experimental data
2. Compute >300 Descriptors
   - Constitutional
   - Topological
   - H bonding
   - Ionization
   - Electrotopological
   - Charge/reactivity
3. Remove low variance and correlated descriptors
4. Test Set Selection
   1) Kohonen map
   2) Every nth
   3) Random
   4) K-means
   5) Manual
5. Apply model to test set
6. Training Set
7. ANNE Training
8. Grid of Network Ensembles (X descriptors by Y neurons)
9. Select best model
10. ANNE
11. No. of neurons and descriptors
   - Create models with different architectures
   - Sensitivity analysis
   - Which descriptors create the best model

---

Test Set Selection:
1) Kohonen map
2) Every nth
3) Random
4) K-means
5) Manual

Train Set:
- ANNE Training
- Grid of Network Ensembles (X descriptors by Y neurons)
- Select best model
- ANNE

---

Simulations Plus, Inc.
Classification Neural Network

Descriptors: \( X_i \)
Normalized to range 0.0-1.0

\[
f_j = \tanh\left( \sum_{i} w_{ij} x_i - t_j \right)
\]

Weights: \( f_1, f_2, f_3, \ldots, f_n \)

Neurons: \( a_1, a_2, a_3 \)

Output:

\[
g = \log \left( \sum_{j} a_j f_j - b \right)
\]

Yes \quad Predicted “Positive”

No \quad Predicted “Negative”

Weights and threshold adjusted iteratively to optimize model performance on the training set.

\[
Obj = \sum_{k=1}^{n} w_0 (1 - c(k))(g(k))^2 + w_1 c(k)(1 - g(k))^2
\]

where \( c(k) \) is 0 if observation \( k \) is in the negative class and 1 if observation \( k \) is in the positive class.
Stopping Early to Avoid Overtraining

- Training set
- Verification set

Local minima for the verification set

Stop when the objective function for the verification set is higher than last minimum for $m$ consecutive iterations.
The logP Data Set

threshold: logP = 2
(~50:50 split)

N = 12580

“negatives”

“positives”
Outline

• Motivation
• Model building in ADMET Modeler™
  • Binomial, beta and beta-binomial distributions for modeling error distributions
• Fitting an uncertainty profile to the training pool
• Applying an uncertainty profile to the test set
• Using averaging instead of voting
A Shift in Model Perspective

ANNE architecture: 36 inputs x 5 neurons x 33 networks
629 compound random training pool (train + verify)

Negatives & Positives

Predictions & Errors

A Shift in Model Perspective

ANNE architecture: 36 inputs x 5 neurons x 33 networks
629 compound random training pool (train + verify)
The Binomial Approach

• If the $K$ network predictions are independent of one another, the errors should follow a binomial distribution across the number of positive votes $k$:

$$Binom(k|K,p) = \binom{K}{k} p^k (1 - p)^{K-k}$$

• That grossly underestimates the spread in errors, because the networks in an ANN ensemble are not independent.
  – in addition, if they were independent the overall error rate would be expected to go down as the square root of the number of networks in the ensemble; that does not generally happen

• Tried estimating an effective number of degrees of freedom
  – that did not work very well either

• What’s the alternative?
Enter the Beta Binomial

• The beta binomial is a variant of the “usual” binomial distribution in which the probability of success $p$ varies:

$$p \sim \text{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

(note: $\Gamma(n) = (n-1)!$ and $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$)

• It is used in the biometrics literature for series of events which are not independent of each other (e.g., accumulated mutations)

$$\text{BB}(k|K, \alpha, \beta) = \binom{K}{k} \frac{\text{B}(k + \alpha, K - k + \beta)}{\text{B}(\alpha, \beta)}$$
Meet the Beta Distributions

Beta Distributions

- $\alpha = 0.3, \beta = 0.4$
- $\alpha = 6, \beta = 3$

Beta Binomial Distributions

- $\alpha = 0.3, \beta = 0.4$
- $\alpha = 1, \beta = 1$ (random)
- $K = 33$
Outline

• Motivation
• Model building in ADMET Modeler™
• Binomial, beta and beta-binomial distributions for modeling error distributions
• Fitting an uncertainty profile to the training pool
• Applying an uncertainty profile to the test set
• Using averaging instead of voting
Fitting Training Pool Uncertainty

1. Build an ensemble of $K$ networks, each with its own threshold
   - maximizing the Youden index $J = (\text{sensitivity} + \text{specificity}) - 1$
2. Tally the number of positive votes $k$ for each prediction
3. Set the voting threshold to $k^* = 0.5 K$
4. Classify negatives with $k > k^*$ as errors
5. Classify positives with $k < k^*$ as errors
6. Add a continuity correction to the count for each tally
   - necessary to mitigate problems with undersampling
   - add 1 for predictions and 0.5 for errors
7. Fit the prediction distribution to a beta binomial $\varphi(k)$
8. Fit the error distribution to a beta binomial $\varepsilon(k)$
9. Estimate the uncertainty distribution by $u(k) = FP \varepsilon(k)/\varphi(k)$, where $FP$ is the overall false positive rate for the training pool
10. Calculate the estimated confidence as $1 - u(k)$

fit to cumulative distributions
Training Pool Uncertainty

ANNE architecture: 36 inputs x 5 neurons x 33 networks
629-compound random training pool (train + verify)

Count

training pool predictions

Tally of positive votes

errors

Observed

Fitted

errors

Uncertainty

Tally of positive votes

PREDICTIONS
ERRORS

OBSERVED
FITTER

©Simulations Plus, Inc., 2013
All rights reserved
Outline

- Motivation
- Model building in ADMET Modeler™
- Binomial, beta and beta-binomial distributions for modeling error distributions
- Fitting an uncertainty profile to the training pool
- Applying an uncertainty profile to the test set
- Using averaging instead of voting
It May Look Ugly, But Is It Predictive?

ANNE architecture: 36 inputs x 5 neurons x 33 networks
11951-compound random test set (95%)

threshold = 16.5 votes

YES!
Why Does It Work?

• Continuity corrections suppress noise due to sparse sampling in the center of the distribution and force a limiting uncertainty of 0.5, which is the expected optimal value at the threshold.

• Fitting to the cumulative distribution functions ensures that the high and low ends of the tally range – which are typically well-populated – dominate the curvatures.
Other Examples

• Ames mutagenicity
  – 6471 compounds used with some curation of structures
  – 2983 compounds classed as “active”

• CYP2D6 inhibition
  – NCGC luciferase-based qHTS screen: PubChem: AID 1851
  – data on 5959 compounds used with some curation of structures
  – 2806 compounds with AC50 < 10μM classed as “positive”
Ames Mutagenicity

ANNE architecture: 16 inputs x 3 neurons x 33 networks
5872-compound random test set (90%)

log scale

Count

Tally of positive "votes"

PREDICTIONS

ERRORS

Uncertainty

Tally of positive "votes"

observed test set uncertainty

fitted training pool uncertainty

©Simulations Plus, Inc., 2013
All rights reserved
CYP2D6 Inhibition

ANNE architecture: 25 inputs x 3 neurons x 33 networks
5359-compound random test set (90%)
Outline

• Motivation
• Model building in ADMET Modeler™
• Binomial, beta and beta-binomial distributions for modeling error distributions
• Fitting an uncertainty profile to the training pool
• Applying an uncertainty profile to the test set
• Using averaging instead of voting
Application to Averaged Outputs

Process parallels that for the voting method except that:

• Outputs are averaged across the networks in the ensemble and the average outputs $\bar{x}$ replace vote tallies
  – average output is a real number between 0 and 1

• The initial threshold is set to the value that provides the highest Youden index

• Predictions are collapsed inwards before fitting so as to remove “tails” at the high and low ends of the range

• The errors and prediction distributions are fit to beta functions rather than to beta binomials
logP Classification (Averaging)

ANNE architecture: 15 inputs x 3 neurons x 33 networks
11951-compound random test set (95%)

Development is ongoing…
Take-Home Messages

- Fitting ensemble misclassifications to a binomial distribution across vote tallies is unlikely to work well.
- ANNE prediction and error profiles follow beta binomial distributions.
- The uncertainty of a prospective classification can be estimated from the results for the training pool.
- Prospective uncertainty estimates are reliable for ANNEs built using early stopping to avoid overfitting.
- The method used for ensemble voting can be applied to ensemble averaging by fitting to a beta distribution instead of to a beta binomial distribution.
Confidence Estimates in ADMET Predictor™ 6.5

<table>
<thead>
<tr>
<th>Number</th>
<th>Compound Structure</th>
<th>TOX_MUT_90</th>
<th>TOX_MUT_m90</th>
<th>TOX_MUT_100</th>
<th>TOX_MUT_n100</th>
<th>TOX_MUT_102wp2</th>
<th>TOX_MUT_n102wp2</th>
<th>TOX_MUT_104wp2</th>
</tr>
</thead>
<tbody>
<tr>
<td>4393</td>
<td><img src="image" alt="Image" /></td>
<td>Negative (99%)</td>
<td>Negative (99%)</td>
<td>Negative (99%)</td>
<td>Negative</td>
<td>Negative (99%)</td>
<td>Negative (99%)</td>
<td>Negative</td>
</tr>
<tr>
<td>4436</td>
<td><img src="image" alt="Image" /></td>
<td>Negative (93%)</td>
<td>Positive (82%)</td>
<td>Negative (99%)</td>
<td>Positive</td>
<td>Negative (99%)</td>
<td>Negative (97%)</td>
<td>Negative</td>
</tr>
<tr>
<td>62</td>
<td><img src="image" alt="Image" /></td>
<td>Negative (99%)</td>
<td>Positive (50%)</td>
<td>Negative (99%)</td>
<td>Negative</td>
<td>Negative (99%)</td>
<td>Negative (99%)</td>
<td>Negative</td>
</tr>
<tr>
<td>596</td>
<td><img src="image" alt="Image" /></td>
<td>Positive (87%)</td>
<td>Positive (53%)</td>
<td>Positive (51%)</td>
<td>Positive</td>
<td>Negative (99%)</td>
<td>Negative (96%)</td>
<td>Negative</td>
</tr>
</tbody>
</table>
Acknowledgements

co-authors

- Wenkel Liang
- Marvin Waldman
- Robert Fraczkiewicz
- Michael Lawless
- Adam Lee
- Jinhua Zhang
- Michael Bolger
- Walter S. Woltosz

Thank you