How Well Can a QSAR Model Handle New Datasets?

Rajarshi Guha and Peter C. Jurs
Department of Chemistry
The Pennsylvania State University

1

Validating QSAR Models

- QSAR models are built using a training subset
- They are validated using a prediction subset
- The model is not usually tested on unknown data
- Predictive quality is generally indicated by prediction set statistics
- Examples of statistics include:
  - R^2
  - RMSE
  - Cross - validation

2

What About New Data?

- Given a model we would like to know how it will perform when faced with new data
- Trivial solution – run the data through the model
- But can we determine how well the model will perform without running the model?
- The question is: *how can we quantify predictive ability when faced with new data*

3

Two Approaches

Similarity

- Choose a measure of model quality. Restricted choice if we want to maintain generality
- Attempt to correlate this with a similarity measure
- A variety of similarity metrics are available
- Use an enrichment scheme, rather than direct comparison to obtain better results

Classification

- Directly predict whether a compound will be well predicted or not
- Involves arbitrary class assignments to the training set
- Wide variety of classification algorithms
- Allows us to get a probability associated with the class prediction

4

Datasets

Artemisinin Analogs	DIPP Dataset
• 179 molecules	• 277 molecules
• cutoff = 1.0	• cutoff = 1.0
• 131 *good* molecules	• 213 *good* molecules
• 46 *bad* molecules	• 64 *bad* molecules

Linear Models

Artemisinin Analogs	DIPP Dataset
• 4 descriptor model	• 4 descriptor model
• N7CH, NSB,WTPT,MDE14	• FPSA, FNSA, RNCG, RPCS
• R^2 = 0.70 RMSE = 0.87	• R^2 = 0.99 RMSE = 7.22

5

Sphere Exclusion & Similarity

- Each PSET point is the center of a *k*-D sphere
- Radius of the sphere is the radius of a *k*-D sphere of unit volume based on the total volume of the dataset
- A Monte Carlo method is used to correct the total volume for uneven distribution of the points

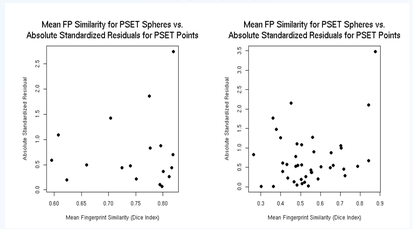
6

Using the Spheres

- Consider TSET density within a sphere
  - Number of TSET points in a sphere divided by total number of TSET points
- Calculate similarity metric with TSET points in a sphere
  - Fingerprint similarity and atom pair similarity
- Correlate similarity with model quality

7

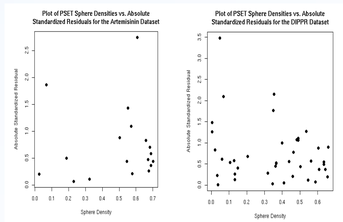
Fingerprint Similarity & Residuals



Artemisinin DIPPR

8

Sphere Density & Residuals



Artemisinin DIPPR

9

Classification Choices

- How do we decide on *good* or *bad*?
  - Regression diagnostics
  - Standardized residuals
- How do we build a classifier?
  - Linear methods (LDA, PLS)
  - Non-linear methods (decision trees, CNN)

10

Classification of Residuals

- The goal is to decide whether a molecule will be well predicted or not
  - *Well predicted implies a low residual*
- Build a classifier using the models' training set and model descriptors
- Training set residuals are arbitrarily assigned as good or bad
- The new dataset is then run through the classifier to obtain a predicted class

11

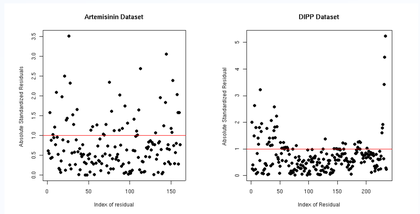
Assigning Classes to Residuals

- The assignment of classes to the training set is arbitrary
- We restricted ourselves to 2 classes
- A split value, **s**, chosen by the user, so that

$$abs(\text{Std. Res.}) \geq s \Rightarrow \text{bad}$$
$$abs(\text{Std. Res.}) < s \Rightarrow \text{good}$$

12

The Distribution of Residuals In The Two Classes



13

Handling Unbalanced Classes

- The class assignments can lead to very skewed classes
- This can be alleviated by
  - Oversampling the minority class
  - Undersampling the majority class
  - Extending the dataset using convex pseudo data

14

CNN Classifier Performance

Artemisinin

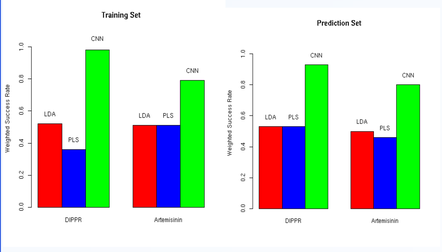
	Training Set		Prediction Set		
	Predicted			*Predicted*	
Actual	Bad	Good	*Actual*	Bad	Good
Bad	54	0	Bad	8	2
Good	5	176	Good	2	30

DIPPR

	Training Set		Prediction Set		
	Predicted			*Predicted*	
Actual	Bad	Good	*Actual*	Bad	Good
Bad	34	8	Bad	3	1
Good	46	73	Good	4	10

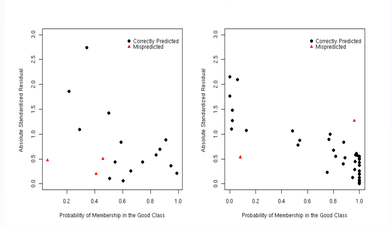
15

Weighted Classification Rates



16

Classification Diagnostic Plots



Artemisinin DIPPR

17

Why Use The Classification Methodology?

- It only considers residuals and so it can be applied to any type of quantitative model, linear or non-linear
- Does not require the original model
- In the absence of confidence scores for a given model, this method can provide a confidence measure for predictions

18

Further Work

- More than two classes
  - Requires a large dataset
- Automated class assignments
  - Use regression diagnostics
  - Might lead to a loss of generality
- Bayesian classification approach
  - Build a prior probability distribution and determine probability of class membership for new compounds by sampling this distribution