Making the Most of Predictive Models

Rajarshi Guha

School of Informatics
Indiana University

28th February, 2007
CUP 8, Santa Fe

Outline

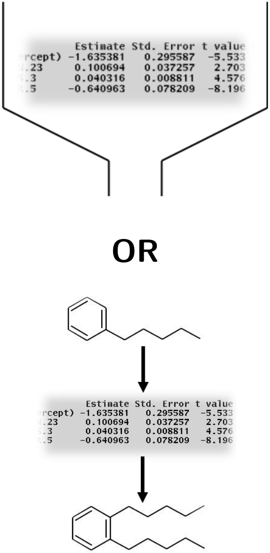
Outline

All Models Predict Something . . .

- Physics based models
  - Docking
  - Force fields
  - MD
- Statistical/ML models
  - Indirect description of the physical situation
  - Not always clear as to how a prediction is made

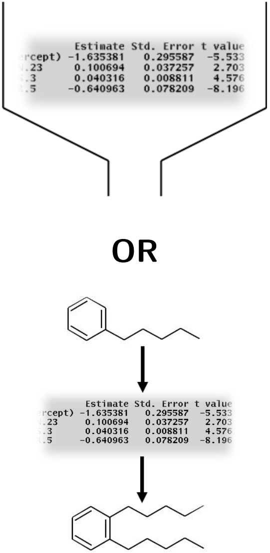
The Scope of Predictive Models

- Predictive models can be used for
  - filtering
  - analysis
- We can use predictive models in
  - chemometrics
  - bioinformatics
  - QSAR
  - ...
- What do we look for in a model?
  - Validity
  - Accuracy
  - Applicability
  - Interpretability



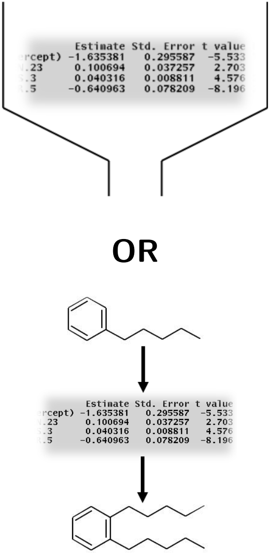
The Scope of Predictive Models

- Predictive models can be used for
  - filtering
  - analysis
- We can use predictive models in
  - chemometrics
  - bioinformatics
  - QSAR
  - . . .
- What do we look for in a model?
  - Validity
  - Accuracy
  - Applicability
  - Interpretability



The Scope of Predictive Models

- Predictive models can be used for
  - filtering
  - analysis
- We can use predictive models in
  - chemometrics
  - bioinformatics
  - QSAR
  - ...
- What do we look for in a model?
  - Validity
  - Accuracy
  - Applicability
  - Interpretability



Why Interpret a Model?

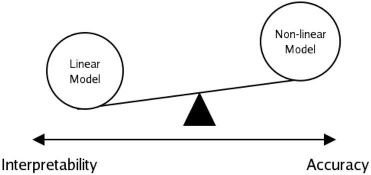
- A model encodes relationships between features and the property
- Understanding these relationships allows us to
  - understand why a molecules is active or not
  - suggest structural modifications
  - explain anomalous observations

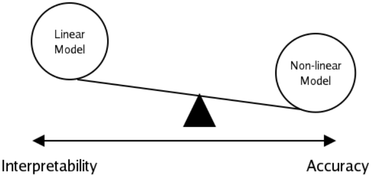
How Much Detail Can We Extract?

- We can look at a model very broadly
  - Which descriptors are important to it's predictive ability?
- We can consider a more detailed analysis
  - What is the effect of descriptor X_i on $\hat{Y}$
  - Which observations highlight this relationship?
- Depends on how much effort you want to put in

The Accuracy - Interpretability Tradeoff

- OLS models are generally easier to interpret but not always accurate
- Neural networks give better accuracy, but are black boxes
- Some lie in between, such as random forests





Outline

Model Types That We Can Interpret

- Depends on the
  - Level of interpretation desired
  - Nature of the problem (classification and regression)
- Some models are interpretable by design
  - Decision trees
  - Bayesian networks
- Other require an interpretation protocol
  - Linear regression
  - Random forests
  - Neural networks
  - Support vector machines

Guha, R.; Jurs, P.C.,*J. Chem. Inf. Comput. Sci.*, **2004**, *44*, 2179–2189

Guha R. et al., *J. Chem. Inf. Model.*, **2005**, *46*, 321–333

Do, T.N.; Poulet, F., Enhancing SVM with Visualization in *Discovery Science*, Springer, **2004**, pp. 183–194

Linear Regression Interpretations

$$pK_a = -37.54 Q_{\sigma,o} + 12.27 A_{access,o} + 0.11 \chi_{\pi,\alpha c} - 1.02 \alpha_o - 1.89 I_{amino} + 19.10$$

- Simply looking at the magnitude of the coefficients describes which descriptors are playing an important role
- Signs of the coefficients indicate the effect of the descriptor on the predicted property
- Interpretation is still quite broad
  - We'd like to see more detailed SAR's applied to individual molecules

Zhang, J. et al., *J. Chem. Inf. Model*, **2006**, *46*, 2256–2266

Linear Regression Interpretations via PLS

- PLS overview
  - Creates a model with *latent variables*
  - Latent variables (components) are linear combinations of the original variables (X's)
  - Each latent variable is used to predict a *pseudo* dependent variable (Y's)
- Interpretation
  - The linear model is subjected to PLS analysis
  - This also *validates* the model
  - Choose the number of components to use
  - Interpretation uses the X-weights, X-scores & Y-scores

Stanton D.T.; *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 1423-1433

Linear Regression Interpretations via PLS

- Choosing components
  - Q^2 allows us to choose how many components
  - For a valid model cumulative variance should be 1
- Descriptor weights
  - Descriptors are ranked by their weights
  - Sign of weight indicates how the descriptor correlates to predicted activity

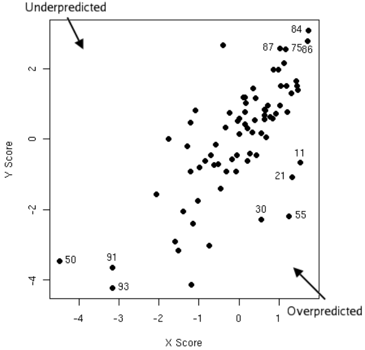
	X variance	R^2	Q^2
C1	0.51	0.52	0.45
C2	0.78	0.60	0.56
C3	1.00	0.61	0.56

Desc	C1	C2	C3
MDEN-23	-0.16	0.93	0.30
RNHS-3	0.55	-0.17	0.81
SURR-5	-0.82	-0.29	0.48

Guha, R.; Jurs, P.C., *J. Chem. Inf. Comput. Sci.*, **2004**, *44*, 2179–2189

Linear Regression Interpretations via PLS

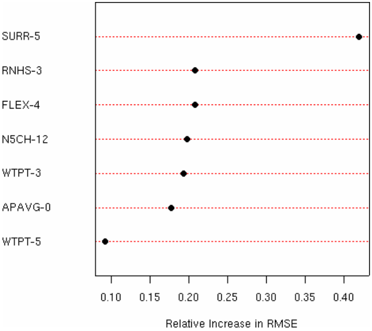
- Component 1
  - SURR-5 is most weighted
  - Low values of SURR-5 $\Rightarrow$ high values of predicted activity
- Interpretation
  - Active compounds have high absolute values of SURR-5
  - Indicates large hydrophobic surface area
  - Consistent with cell based assay which depends on cell membrane transport



	MDEN-23	RNHS-3	SURR-5
C1	-0.16	0.55	-0.82

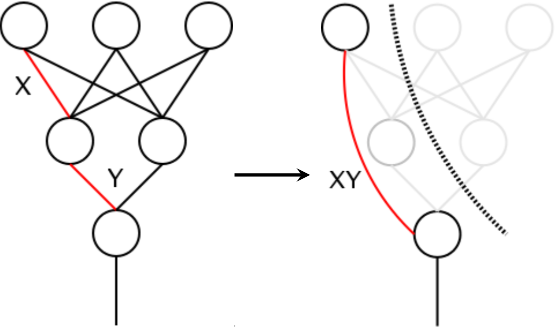
Random Forest Interpretations

- A RF provides a measure of descriptor importance
  - Utilizes the whole descriptor pool and ranks the descriptors
  - Based on randomization
- We can also get more information on individual trees
  - Find the most important trees
  - Consider a tree-space and find clusters of trees



Chipman, H.A. et al., Making Sense of a Forest of Trees in *Proc. 30th Symposium on the Interface*, Weisberg, S. Ed

Neural Network Interpretations



Effective Weight Matrix

	Hidden Neuron	
Descriptor	1	2
Desc 1	52.41	29.30
Desc 2	37.65	22.14
Desc 3	−10.50	−16.85

Linearizes the network and consequently looses some details of the encoded SAR's

Guha R. et al., *J. Chem. Inf. Model.*, **2005**, *46*, 321–333

Neural Network Interpretations

- The most weighted descriptors are very similar to those in the OLS model
- The signs of the effective weights match those from the OLS model as well as chemical reasoning

	Hidden Neuron			
Descriptor	1	3	2	4
PNSA-3	-1.80	**-6.57**	0.39	-1.43
RSHM-1	**4.03**	6.15	1.50	1.01
V4P-5	**9.45**	2.15	**3.24**	0.60
S4PC-12	3.36	2.73	**1.99**	0.56
MW	3.94	**8.42**	1.94	0.76
WTPT-2	1.71	2.61	1.17	-0.13
DPHS-1	0.66	0.44	0.33	1.65
SCV	0.52	0.33	0.13	0.01

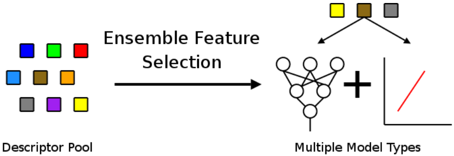
Size effects, higher MW

H-bonding, HSA, polar surface area

Stepping Back . . .

- We've been focusing on single model types
- Different models for different purposes
- Descriptors are optimal for a specific model
- Are we sure that the different models encode the same SAR's?

Getting the Best of Both Worlds



- Why not *force* multiple models to have the same descriptors?
  - The descriptor set will not be optimal for either model
  - Degradation in accuracy
- Is it really that bad?

Eating Our Cake . . .

- Ensemble feature selection
  - Select a descriptor subset that is *simultaneously* optimal for two different model types
- Allows us to build an OLS model and a CNN model using the same set of descriptors
- We use a genetic algorithm, where the objective function is of the form

$$RMSE_{OLS} + RMSE_{CNN}$$

We have one model for interpretability and one for accuracy, but they should now incorporate the same SAR's

Dutta, D. et al., *J. Chem. Inf. Model.*, submitted

Eating Our Cake ...

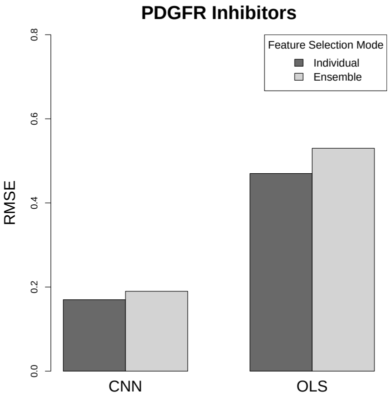
- Ensemble feature selection
  - Select a descriptor subset that is *simultaneously* optimal for two different model types
- Allows us to build an OLS model and a CNN model using the same set of descriptors
- We use a genetic algorithm, where the objective function is of the form

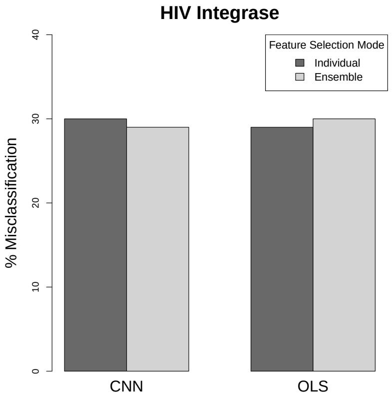
$$RMSE_{OLS} + RMSE_{CNN}$$

We have one model for interpretability and one for accuracy, but they should now incorporate the same SAR's

Dutta, D. et al., *J. Chem. Inf. Model.*, submitted

. . . and Having it Too





Selecting for Interpretability

- Uptil now we've considered the models themselves
- But it's the descriptors we interpret
- Can we build models out of interpretable descriptors?
  - Design descriptors so that they have physical meaning
  - Exclude uninterpretable descriptors from the pool
  - Add semantic annotation to descriptors and modify feature selection algorithms to take this into account

Outline

Summary

- Using models just for prediction is fine, but there's lots of extra information we can extract from them
- The extent of interpretability is guided by the nature of the problem, the choice of model and descriptors
- Interpretation of 2D-QSAR models allows us to get a little closer to the real physical problem

PLS Interpretation - Understanding Outliers

- Compound 55 is mispredicted by each component
- It is also an outlier in both linear & CNN models
- Has high absolute value of SURR-5 but low measured activity