QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP MODELING USING LEO BREIMAN'S RANDOM FOREST

Christopher Tong, Vladimir Svetnik, and Andy Liaw
Biometrics Research, Merck Research Laboratories
RY33-300, P.O. Box 2000, Rahway, NJ 07065
{christopher_tong, vladimir_svetnik, andy_liaw}@merck.com

Key Words: Random Forest, ensemble learning, classification, regression, QSAR, drug discovery and development.

1. Introduction

A common problem in drug discovery and development is quantitative structure-activity relationship (QSAR) modeling [Eki00, Haw01, Liv00, Sto93]. The idea is to use a quantitative description of a chemical compound's structure and/or properties to predict the compound's biological activity. In other words, we wish to model the function,

$$activity = f(structure, properties). \qquad (1)$$

The activity may be a continuous or categorical variable, so that QSAR modeling is considered to consist of regression and classification problems. The structure of the compound is often described by a set of topological descriptors, such as atom pairs [Car85] or topological torsions [Nil87]. The number of descriptors, p, usually exceeds the number of samples, n, so that the traditional statistical methods (multiple linear regression [MLR], linear discriminant analysis [LDA], and k-nearest neighbors [kNN]) cannot be used reliably without a sophisticated variable selection filter, such as a genetic algorithm. This approach is indeed taken by some investigators. Other approaches found in the literature include Decision Tree (recursive partitioning [RP]), Partial Least Squares (PLS), artificial neural networks (ANN), and support vector machines (SVM). Examples of such studies include [Bak00, Don02, Kau01, Rus99].

Unfortunately, each of these approaches suffers from limitations. For instance, simple methods like MLR, LDA, and PLS are not flexible enough, since they lack interactions and the ability to model multiple mechanisms of action. Other methods like ANN and SVM are too flexible, requiring intensive training and parameter tuning. Only the Decision Tree is relatively free of all these limitations; however, it suffers from low accuracy. Ensemble learning methods applied to Trees have recently been introduced to increase their accuracy [Die02]. Examples include bagging [Bre96], boosting [Fre97], Random Forest [Bre01], and Decision Forest [Ton03]; there are numerous others. In this talk we focus on the use of Random Forest for QSAR modeling.

2. Random Forest

Like bagging, Random Forest is an ensemble of unpruned trees. Each tree is trained on a bootstrap sample of the training data. Random Forest differs from bagging in that at each node, the algorithm considers as splitting candidates a random sample of the variables instead of all the variables. The size of the variable subset is a fixed value, $mtry$, with default value $\sqrt{p}$ for classification and $p/3$ for regression. The idea is to maintain the "strength" of the trees while reducing their correlation. Breiman [Bre01] has shown that an upper bound on the generalization error of Random Forest is given by $r(1-s^2)/s^2$, where r is a measure of the correlation between the trees, and s is a measure of their strength (see [Bre01] for the details). Since the unpruned trees are low-bias, high variance models, averaging over an ensemble of trees reduces variance while keeping low bias. This can be demonstrated explicitly by examining the bias-variance decomposition behavior of Random Forest [Sve03b]. It is also thought that an ensemble of trees mitigates the semi-artificiality of the tree structure (hyper-rectangular partition of the descriptor space) and the greediness of the tree-growing algorithm, which are arguably the two drawbacks of the Tree approach.

Random Forest also provides additional features that increase its utility for QSAR modeling:

1. Built-in error estimation, using out-of-bag predictions;
2. A measure of variable importance; and
3. A measure of intrinsic proximity between two compounds.

These features are discussed and analyzed further in [Bre01, Sve03b].

3. QSAR models of P-gp transport

P-glycoprotein (P-gp) is a drug transport protein that lives on the cell membrane, protecting the cell from xenobiotics [Sto02]. A publicly available data set of P-gp transport activities for 186 drug compounds is provided by Penzotti *et al.* of Deltagen [Pen02]. The response variable is a classification of the drug as a substrate or a non-substrate of P-gp. We generated a set of 1522 binary atom pair descriptors for these compounds and performed two different performance assessment procedures. The first was to imitate the original authors' assessment procedure for their own QSAR model. Their model was based on sophisticated 3D pharmacophores [Pen02]; we will call it the Deltagen model. They randomly split the data into a training set and test set, and reported the accuracy rate on the test set (TS). We developed QSAR models based on Random Forest (RF), single tree (RP), and PLS for comparison, and computed their accuracies on the same training/test split of the data. We also calculated median accuracy rates of the latter three models based on 50 replications of 5-fold cross-validation (CV). The accuracy results are shown in the following table.

Assessment Procedure	RF	RP	PLS	Deltagen
TS	0.70	0.70	0.68	0.63
CV	0.806	0.712	0.769	N/A

Using the single test set (TS) assessment procedure, we see that all three of the methods we tried perform equivalently, and they outperform the Deltagen model. However, in cross-validation (CV), it is seen that the Random Forest outperforms the single tree (RP) and does just as well as PLS.

We examined five other QSAR data sets, including regression examples, and showed that Random Forest is consistently among the top performers in terms of accuracy; these results will be reported elsewhere [Sve03b]. Similar conclusions can be drawn from the benchmarking experiments of Meyer *et al.* [Mey03]. (Notably, they also examined another well-known tree ensemble method, MART, a type of boosting [Fri01].) The bottom line is that over a range of diverse data sets, Random Forest has a performance as good as or nearly as good as the best-performing algorithms.

4. Parameter tuning and variable selection

Random Forest has a parameter, *mtry*, that in principle could be considered a tuning parameter. Also, the removal of irrelevant variables may affect the performance of Random Forest. However, we argue that neither of these issues seems to be as critical to performance as they would be in most other machine learning methods. The argument is clearest with regard to variable reduction. Trees are generally resistant to the presence of irrelevant variables, since embedded variable selection is intrinsic to the tree growing process [Guy03]. Ensembles of trees should be even more capable of resisting the influence of irrelevant variables. Therefore, we do not expect Random Forest to gain much in accuracy performance if variable reduction is implemented. As an illustration, again consider the p-glycoprotein transport data. We implemented the following variable reduction algorithm:

1. Partition the data for 5-fold cross-validation.
2. On each CV training set, train a model on all variables and use the variable importance measure to rank them. Record the CV test set predictions.
3. Use the variable ranking to remove the least important half of the variables and retrain the model, predicting the CV test set. Repeat removal of half of the variables until there are 2 left.
4. Aggregate results from all 5 CV partitions and compute the error rate (or MSE) at each step of halving.
5. Replicate steps (1)-(4) 50 times to "smooth out" the variability.

The median error rates for the 50 replications, with medians connected by line segments, is shown in Fig. 1, for various choices of *mtry*. The cases of *mtry* equaling p (equivalent to bagging), $p/2$, $p/4$, and the default $\sqrt{p}$ are considered.

The plot shows that the default *mtry* performs the best, but the other choices are only a few percent worse, still competitive with PLS. Also, the performance remains about the same as irrelevant variables are removed, until you reach 191 variables. Further removal of variables will degrade the prediction performance.

Although the robustness of Random Forest's performance to the presence of irrelevant variables is to be expected, its robustness to changes in *mtry* is a pleasant surprise; users should always investigate this to make sure it will be the case for their data. A more in-depth discussion of these issues can be found in [Sve03a].

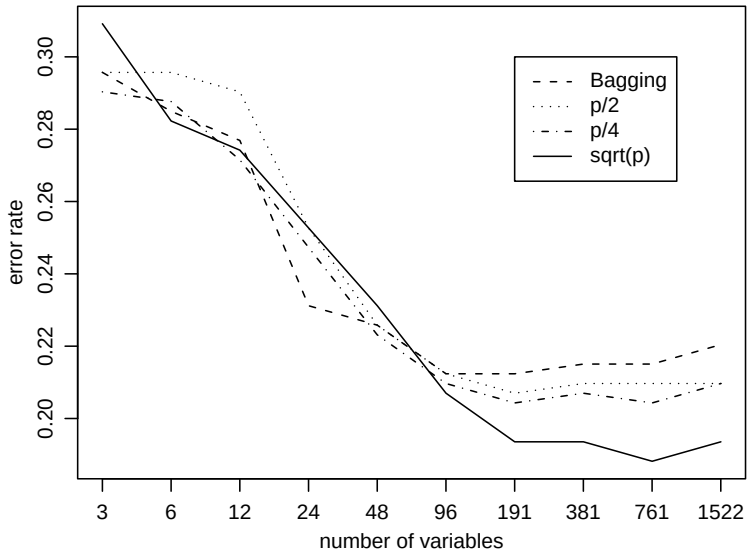

Figure 1: Median CV test error rates at each step of halving the important variables, using different *mtry* functions, for the P-gp data. Line segments connect the medians of 20 5-fold CV error rates.

5. Random Forest software

Open source software for Random Forest is publicly available. The Fortran code for the Random Forest software, written by Leo Breiman and Adele Cutler, is available at `http://www.stat.berkeley.edu/users/breiman/rf.html` and an R interface for it by Andy Liaw and Matt Wiener [Lia02] can be found at `http://cran.us.r-project.org/` by looking for the `randomForest` package. A Matlab interface by Ting Wang is currently under development at Merck; please contact us for further information on it.

6. Conclusion

As discussed in the Introduction, the Decision Tree has the "right" combination of features to make it appealing for QSAR modeling, but it suffers from low predictive accuracy. In this talk, we demonstrate on the P-gp data set that Random Forest outperforms a single tree and can perform as well as other methods like PLS. Random Forest performs well "off the shelf", apparently not requiring much parameter tuning or variable selection (although in serious research, both issues should always be investigated). Open source software is freely available, so we look forward to hearing from other researchers about their successes (or otherwise) with using Random Forest for challenging classification and regression problems.

7. Acknowledgments

We are grateful to Leo Breiman for extensive consultation on Random Forest. We also thank Chris Culberson, Bob Sheridan, and Brad Feuston for insights on QSAR modeling and descriptor generation; Matt Wiener and Ting Wang for help with software development; Randy Tobias for assistance with SAS PROC PLS; David Meyer for correspondence on his work; Tom Dietterich for discussions on ensemble learning; and Peter Grootenhuis and Michelle Lamb for discussions about the P-gp data set and the Deltagen model.

References

[Bak00] Bakken, G. A., Jurs, P. C., Classification of multidrug-resistance reversal agents using structure-based descriptors and linear discriminant analysis. *J. Med. Chem.* **43** (2000) 4534–4541

[Bre96] Breiman, L.: Bagging predictors. *Machine Learning* **26** (1996) 123–140

[Bre01] Breiman, L.: Random Forests. *Machine Learning* **45** (2001) 5–32

[Car85] Carhart, R. E., Smith, D. H., Venkataraghavan, R.: Atom pairs as molecular features in structure-activity studies: definitions and applications. *J. Chem. Inf. Comput. Sci.* **25** (1985) 64–73

[Die02] Dietterich, T. G.: Ensemble learning. In *The Handbook of Brain Theory and Neural Networks*, second edition, edited by M. A. Arbib. (2002) Cambridge, MA: The MIT Press, pp. 405–408.

[Don02] Doniger, S., Hofmann, T., Yeh, J.: Predicting CNS permeability of drug molecules: comparison of neural network and support vector machine algorithms. *J. Comput. Biol.* **9** (2002) 849–864

[Eki00] Ekins, S. *et al.*: Progress in predicting human ADME parameters in silico. *J. Pharmac. Toxic. Meth.* **44** (2000) 251–272

[Fre97] Freund, Y., Schapire, R.: A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.* **55** (1997) 119–139

[Fri01] Friedman, J. H.: Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29** (2001) 1189–1202; Stochastic gradient boosting. *Comp. Stat. Data Anal.* **38** (2002) 367–378

[Guy03] Guyon, I., Elisseef, A.: An introduction to variable and feature selection. *J. Machine Learning Res.* **3** (2003) 1157–1182

[Haw01] Hawkins, D. M., Basak, S. C., Shi, X.: QSAR with few compounds and many features. *J. Chem. Inf. Comput. Sci.* **41** (2001) 663–670

[Kau01] Kauffman, G. W., Jurs, P. C.: QSAR and k-nearest neighbor classification analysis of selective cyclooxygenase-2 inhibitors using topologically-based numerical descriptors. *J. Chem. Inf. Comput. Sci.* **41** (2001) 1553–1560

[Lia02] Liaw, A., Wiener, M.: Classification and regression by randomForest. *R News* **2/3** (2002) 18–22

[Liv00] Livingstone, D. J.: The characterization of chemical structures using molecular properties: a survey. *J. Chem. Inf. Comput. Sci.* **40** (2000) 195–209

[Mey03] Meyer, D., Leisch, F., Hornik, K.: The support vector machine under test. *Neurocomputing*, submitted

[Nil87] Nilakantan, R., Bauman, N., Dixon, J. S., Venkataraghavan, R.: Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. *J. Chem. Inf. Comput. Sci.* **27** (1987) 82–85

[Pen02] Penzotti, J. E., Lamb, M. L., Evensen, E., Grootenhuis, P. D. J.: A computational ensemble pharmacophore model for identifying substrates of p-glycoprotein. *J. Med. Chem.* **45** (2002) 1737–1740

[Rus99] Rusinko, A., *et al.*: Analysis of a large structure/biological activity data set using recursive partitioning. *J. Chem. Inf. Comput. Sci.* **39** (1999) 1017–1026

[Sto93] Stone, M., Jonathan, P.: Statistical thinking and technique for QSAR and related studies. I. General theory. *J. Chemometrics* **7** (1993) 455–475

[Sto02] Stouch, T., Gudmundsson, O.: Progress in understanding the structure-activity relationships of p-glycoprotein. *Adv. Drug Delivery Rev.* **54** (2002) 315–328

[Sve03a] Svetnik, V., Liaw, A., Tong, C.: Variable selection in Random Forest with application to quantitative structure-activity relationship. *Proceedings of the 7th Course on Ensemble Methods for Learning Machines.* 22-28 September 2002, Vietri sul Mare, Salerno, Italy. Springer Lecture Notes in Artificial Intelligence, submitted

[Sve03b] Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., Feuston, B. P.: QSAR modeling using Random Forest, an ensemble learning tool for regression and classification. *J. Chem. Inf. Comput. Sci.*, submitted

[Ton03] Tong, W., Hong, H., Fang, H., Xie, Q., Perkins, R.: Decision Forest: Combining the predictions of multiple independent decision tree models. *J. Chem. Inf. Comput. Sci.* **43** (2003) 525–531