Decision Forest: Combining the Predictions of Multiple Independent Decision Tree Models

Weida Tong,*,† Huixiao Hong,‡ Hong Fang,‡ Qian Xie,‡ and Roger Perkins‡

Center for Toxicoinformatics, Division of Biometry and Risk Assessment, National Center for Toxicological Research, Jefferson, Arkansas 72079, and Northrop Grumman Information Technology, Jefferson, Arkansas 72079

Received September 16, 2002

The techniques of combining the results of multiple classification models to produce a single prediction have been investigated for many years. In earlier applications, the multiple models to be combined were developed by altering the training set. The use of these so-called resampling techniques, however, poses the risk of reducing predictivity of the individual models to be combined and/or over fitting the noise in the data, which might result in poorer prediction of the composite model than the individual models. In this paper, we suggest a novel approach, named Decision Forest, that combines multiple Decision Tree models. Each Decision Tree model is developed using a unique set of descriptors. When models of similar predictive quality are combined using the Decision Forest method, quality compared to the individual models is consistently and significantly improved in both training and testing steps. An example will be presented for prediction of binding affinity of 232 chemicals to the estrogen receptor.

INTRODUCTION

The Decision Tree method determines a chemical's activity through a series of rules based on selection of descriptors. These rules are operated by using IF−THEN expressions and displayed as limbs in the form of a *tree* containing, in most cases, only binary branching. For example, a simple rule could be "IF molecular weight > 300, THEN the chemical is active". The rules provide intuitive interpretation of biological questions with respect to the relationship and/or association between descriptors, that is more appealing for some users than a nonlinear "black box" such as an artificial neural network (ANN). One major advantage of Decision Tree is speed of model development and prediction. In the case of the now widespread use of combinatory synthesis in conjunction with high throughput screening (HTS) in drug discovery, Decision Tree offers advantages to quickly process a large volume of data and provide immediate feed back to narrow down the number of chemicals for synthesis and evaluation.[1,2]

The automatic tree construction in Decision Tree dates back to the early 1960s.[3] The Classification and Regression Tree (CART) developed by Breiman et al.[4,5] is widely used in various disciplines. Depending on the nature of the activity data, the tree can be constructed for either regression or classification. Each end node ("leaf of the tree") of a regression tree gives a quantitative prediction, while the classification tree gives categorical predictions. The classification tree is most commonly used in data analysis, where the endpoint is usually binomial (i.e. yes/no or +/−). Since tree-construction methods are recursive in nature, it is also called recursive partitioning (RP) in pattern recognition.

Whether Decision Tree is more accurate than other similar techniques depends on the application domain and the effectiveness of the particular implementation. Lim and Loh[6] compared 22 Decision Tree methods with nine statistical algorithms and two ANN approaches on 32 data sets. They found no statistical difference among the methods evaluated. For classification of estrogen ligands into active and inactive groups, we found that Decision Tree gives comparable results compared to K-Nearest Neighbor (KNN), Soft Independent Modeling of Chemical Analogy (SIMCA), and ANN.[7] It appears that the nature of descriptors used, and more particularly the effectiveness in which they encode the structural features of the molecule related to the activity, is far more critical than the specific method employed.

Evaluating different ways for tree construction and implementation has been a major focus for improving Decision Tree performance. Some representative researches include AID,[3] CHAID,[8] C4.5,[9,10] S-Plus tree,[11] FACT,[12] QUEST,[13] IND,[14] OC1,[15] LMDT,[16] CAL5,[17,18] and T1.[19] Decision Tree methods are also applied in the drug discovery field, such as (1) Statistical Classification of Activities of Molecules (SCAM) developed by Young et al.[1,2] for generation of SAR rules using the binary descriptors in a sequential screening approach; (2) combining RP with simulated annealing (RP/SA) reported by Blower et al.[20] to identify combination of descriptors that give the best tree models; and (3) a novel regression tree based on artificial ant colony systems developed by Izrailev and Agrafiotis.[21]

In this paper, a novel approach is explored that classifies a new chemical by combining the predictions from multiple classification tree models. This method is named Decision Forest, and a model consists of a set of individually trained classification trees that are developed using unique sets of descriptors. Our results suggest that the Decision Forest

* Corresponding author phone: (870)543-7142; fax: (870)543-7662; e-mail: wtong@nctr.fda.gov. Corresponding address: NCTR, 3900 NCTR Road, HFT 20, Jefferson, AR 72079.

† National Center for Toxicological Research.

‡ Northrop Grumman Information Technology.

10.1021/ci020058s
Published on Web 02/04/2003

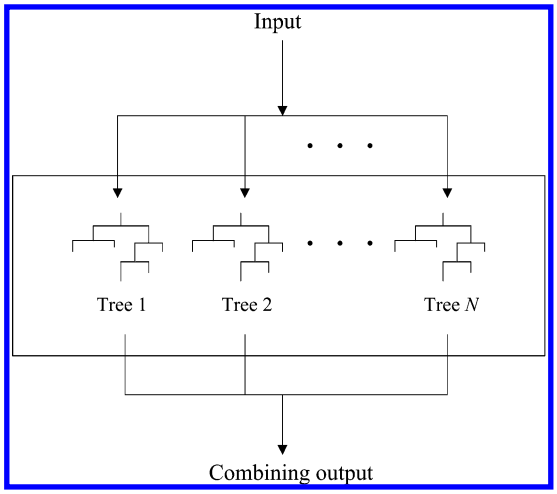

Figure 1. A schematic presentation of combining the predictions of multiple Decision Tree models.

model is consistently superior to any individual trees that are combined to produce the forest in both training and validation steps.

DECISION FOREST

Methodological Consideration. Combining (or ensemble or consensus) forecast is a statistical technique that combines the results of multiple individual models to reach a single prediction.[22] The overall scheme of the technique is shown in Figure 1, where the individual models are normally developed using an ANN[23−25] or Decision Tree.[26,27] A thorough review of this subject can be found in a number of papers.[28−30]

In most cases, individual models are developed using a portion of chemicals randomly selected from the original data set.[31] For example, a data set can be randomly divided into two sets, 2/3 for training and 1/3 for testing. A model developed with the training set will be accepted if it gives satisfactory predictions for the testing set. A set of predictive models is generated by repeating this procedure, and the predictions of these models are then combined when predicting a new chemical. The training set can also be generated using more robust statistical "resampling" approaches, such as Bagging[32] or Boosting.[33]

Bagging is a "bootstrap" ensemble method by which each model is developed on a training set that is generated by randomly selecting chemicals from the original data set.[32] In the selection process, some chemicals may be repeated more than once, while others may be left out so that the training set is the same size as the original data set. In Boosting, the training set for each model is also the same size as the original data set. However, each training set is determined based on the performance of the earlier model(s); chemicals that are incorrectly predicted by the previous model are chosen more often than chemicals that were correctly predicted in the next training set.[33] Boosting, Bagging, and other resampling approaches have all been reported to improve predictive accuracy.

The resampling approaches use only a portion of the data set for constructing the individual models. Since each chemical in a data set encodes some SAR information, reducing the number of chemicals in a training set for model construction will weaken most individual models' predictive accuracy. It follows that reducing the number of chemicals also reduces the improvement in a combining system gained by the resampling approach. Moreover, Freund and Schapire reported that some resampling techniques could be at risk of overfitting the noise in the data, which leads to much worse prediction from multiple models.[33]

The idea of combining multiple models implicitly assumes that one could not identify all aspects of the underlying variable relationship, and thus different models are able to capture it for prediction. Combining several identical models produces no gain. The benefit of combining multiple models can be realized only if individual models give different predictions. An ideal combined system should consist of several accurate models that disagree in prediction as much as possible. Thus, the important aspects of the Decision Forest approach were as follows:

1. Each individual model in Figure 1 is developed using a *distinct* set of descriptors that was explicitly excluded from all other models, thus ensuring each individual model's unique contribution to making prediction.

2. The quality of all models in Decision Forest is *comparable* to ensure that each model significantly contributes to the prediction.

Decision Forest Algorithm. The development of the Decision Forest algorithm consists of the following steps:

1. The algorithm can be initiated with either a predefined *N* to determine the number of models to be combined or a misclassification threshold to set a quality criterion for individual models. The former case is illustrated in this paper.

2. A tree is constructed without pruning. The tree identifies the minimum number of misclassified chemicals (*MIS*) for a given data set. *MIS* then serves as a quality criterion to guide individual tree construction and pruning in the following iterative steps 3−6.

3. A tree is constructed and pruned. The extent of pruning is determined by the *MIS*. The pruned tree assigns a probability (0−1) to each chemical in the data set.

4. The descriptors used in the previous model are removed from the original descriptor pool, and the remaining descriptors are used for the next tree development.

5. Steps 3 and 4 are repeated until no additional model with misclassifications ≤ *MIS* can be developed from the unused portion of the original pool of descriptors.

6. If the total number of models is less than *N*, the *MIS* is increased by 1, and the steps 3−5 are repeated. Otherwise, multiple decisions from individual trees are combined using a linear combination method, where the mean value of the probabilities for all trees is used to determine the classification of a chemical. A chemical with the mean probability larger than 0.5 is designated as active, while a chemical with a mean value less than 0.5 is designated as inactive.

MATERIALS AND METHODS

Tree Development. In the present application, the development of a tree model consists of two steps, tree construction and tree pruning. In the tree construction process, a parent population is split into two children nodes that become parent populations for further splits. The splits are selected to maximally distinguish the response descriptors in the left and right nodes. Splitting continues until chemicals in each

node are either in one activity category or cannot be split further to improve the model. To avoid overfitting the training data, the tree needs to be cut down to a desired size using tree cost-complexity pruning. The method for the tree development is described by Clark and Pregibon[11] as implemented in S-Plus, which is a variant of the CART algorithm. It employs deviance as the splitting criterion. The Decision Forest is written in S language and run in S-Plus software.

Model Assessment. Misclassification and concordance are used to measure model quality. Misclassification is the number of chemicals misclassified in a model, while concordance is the number of correct predictions divided by the total number of predictions.

NCTR Data Set. A large and diverse estrogen data set, called the NCTR data set,[34,35] was used in this study (Table 1). The NCTR data set contains 232 structurally diverse chemicals,[36] of which 131 chemicals exhibit estrogen receptor binding activity,[7] while 101 are inactive[37] in a competitive estrogen receptor binding assay.

Descriptors. More than 250 descriptors for each molecule were generated using Cerius 2 software (Accelrys Inc., San Diego, CA 92121). These descriptors were categorized as (1) conformational, (2) electronic, (3) information content, (4) quantum mechanical, (5) shape related, (6) spatial, (7) thermodynamic, and (8) topological. The descriptors were preprocessed by removing those with no variance across the chemicals. A total of 202 descriptors were used for the final study.

RESULTS

Figure 3 gives a plot of misclassification versus the number of combined decision trees. The number of misclassifications varies inversely with the number of decision trees. The reduction in misclassification is greatest in the first four decision trees combined, where more than 1/2 the misclassifications were eliminated. A decision forest comprising seven trees eliminated about 2/3 of the misclassification of the initial decision tree.

Table 2 provides more detailed results on the decision forest and the decision trees combined. Based on misclassifications, all decision forest combinations perform better than any individual decision tree. Of 202 original descriptors, 88 were ultimately used for the decision forest combining seven decision trees. The progressive decrease in misclassifications as decision trees are successively added to the forest demonstrates how each distinct descriptor set contributes uniquely to the aggregate predictive ability of the forest. Generally, decision trees with fewer "leaves" are expected to perform better because the descriptors are better able to encode the functional dependence of activity on structure. Table 2 also shows the expected trends of both more descriptors and more leaves in the later decision trees as the descriptors that are better able to encode the activity in the previous models are successively removed from the descriptor pool.

Table 3 gives a comparison of decision tree with decision forest as measured by chemicals predicted as active that are actually inactive (false positives) and chemicals predicted as inactive that are actually active (false negatives). The decision tree being compared corresponds to that in the first row of Table 2 that has 17 misclassifications. The Decision Forest being compared in Table 3 corresponds to the bottom row in Table 2 where seven decision trees are combined and for which there are five misclassifications. In the Table 3 comparison, the decision tree utilizes 10 descriptors and produces nine false negatives and eight false positives. In contrast, the Decision Forest utilizes 88 unique descriptors and produces four false negatives and one false positive, a marked improvement in the prediction performance compared to the decision tree. There are 13 chemicals that have contrary activity classification between the decision tree and forest, of which 12 chemicals are correctly predicted by the forest and one is misclassified.

Among the many schemes to combine multiple decision trees, we evaluated linear combination and voting. The voting method uses the majority of votes to classify a chemical. The linear combination method uses the mean of probabilities of the individual decision trees. We found the two methods to produce the same results (results not shown) and chose linear combination because a tie vote is not usable.

Decision Forest assigns a mean probability of the combined trees using the linear combination approach. Figure 4 shows the concordance results of the Decision Forest prediction of the NCTR data set in 10 even intervals between 0 and 1. Analysis shows that the interval 0.7−1.0 has an average concordance of 100% of true positives, and the interval 0.0−0.3 has an average concordance of 98.9% true negatives. The vast majority of misclassifications occur in the 0.3−0.7 probability range where the average concordance is 78%.

A more robust validation of the predictive performance was conducted by dividing the NCTR data set into a training component comprising two-thirds, or 155, of the chemicals and a testing component comprising the remaining 77 chemicals. Both Decision Forest and Decision tree models were constructed for a random selection of the training set and then used to predict the testing set. This was repeated 2000 times to give the concordance results shown in Figure 5. Figure 5 gives on the *Y*-axis the number of times out of 2000 that a model attained the concordance value given on the *X*-axis. The consistently better predictive average concordance of the Decision Forest is readily discernible, as is the narrower distribution for prediction of the training set versus the test set. Both leave-one-out and leave-10-out validation tests were also performed and showed a similar trend (results not shown).

DISCUSSION

We presented a novel combining forecast approach, named Decision Forest that combines predictions of individually trained Decision Trees, each developed using unique descriptors. The method was illustrated by classifying 232 chemicals into estrogen and non-estrogen receptor-binding categories. We demonstrated that Decision Forest yielded better classification and prediction than Decision Tree in both training and validation steps.

A SAR equation can be generalized as Bio $= f(D_1, D_2, ..., D_n)$, where Bio is biological activity data (binomial data in classification) and D_1 to D_n are descriptors. This equation implies that the variance in Bio is explained in a chemistry space defined by the descriptors (D_1 ... D_n). Accordingly,

Table 1. NCTR Data Set, 232 Chemicals with Estrogen Receptor Binding Data

Inactives

1,3-dibenzyltetramethyldisiloxane
1,6-dimethylnaphthalene
1,8-octanediol
2,2′,4,4′-tetrachlorobiphenyl
2,2′-dihydroxy-4-methoxybenzophenone
2,2′-dihydroxybenzophenone
2,3-benzofluorene
2,4,5-T
2,4-D (2,4-dichlorophenoxyacetic acid)
2-chlorophenol
2-ethylphenol
2-furaldehyde
2-hydroxy biphenyl
2-hydroxy-4-methoxybenzophenone
3,3′,4,4′-tetrachlorobiphenyl
4,4′-diaminostilbene
4,4′-dichlorobiphenyl
4,4′-methylenebis(2,6-di-*tert*-butylphenol)
4,4′-methylenebis(*N*,*N*-dimethylaniline)
4,4′-methylenedianiline
4′,6,7-trihydroxy isoflavone
4-amino butylbenzoate
4-aminophenyl ether
6-hydroxy-2′-methoxy-flavone
7-hydroxyflavone
alachlor
aldosterone
aldrin
amaranth
atrazine
benzyl alcohol
bis(2-ethylhexyl)phthalate
bis(2-hydroxyphenyl)methane
bis(*n*-octyl)phthalate
butylbenzylphthalate
caffeine
carbaryl
carbofuran
catechin
chlordane
cholesterol
chrysene
chrysin
cineole
cinnamic acid
corticosterone
dexamethasone
di-2-ethylhexyl adipate
dibenzo-18-crown-6
dieldrin
diethyl phthalate
diisononylphthalate
di-*i*-butyl phthalate (DIBP)
dimethyl phthalate
di-*n*-butyl phthalate (DBuP)
dopamine
endosulfan, technical grade
epitestesterone
ethyl cinnamate
etiocholan-17β-ol-3-one
eugenol
flavanone
flavone
folic acid
genistin
heptachlor
heptaldehyde
hesperetin
hexachlorobenzene
hexyl alcohol
isoeugenol
lindane (gama-HCH)
melatonin
metolachlor
mirex
naringin
n-butylbenzene
nerolidol
o,*p*′-DDD
o,*p*′-DDE
p,*p*′-DDD
p,*p*′-DDE
p,*p*′-DDT
p,*p*′-methoxychlor
p,*p*′-methoxychlor olefin
phenol
progesterone
prometon
quercetin
sec-butylbenzene
simazine
sitosterol
suberic acid
taxifolin
testosterone
thalidomide
trans,*trans*-1,4-diphenyl-1,3-butadiene
trans-4-hydroxystilbene
triphenyl phosphate
vanillin
vinclozolin

Actives

1,3-diphenyltetramethyldisiloxane
16β-hydroxy-16-methyl-3-methyl ether-17β-estradiol
17α-estradiol
17-deoxyestradiol
2,2′,4,4′-tetrahydroxybenzil
2,2′-methylenebis(4-chlorophenol)
2,3,4,5-tetrachloro-4′-biphenylol
2′,4,4′-trihydroxychalcone
2,4′-dichlorobiphenyl
2,5-dichloro-4′-biphenylol
2,6-dimethyl hexestrol
2-chloro-4-biphenylol
2-cholor-4-methyl-phenol
2-ethylhexyl-4-hydroxybenzoate
2-hydroxy-estradiol
2-*sec*-butylphenol
3,3′,5,5′-tetrachloro-4,4′-biphenyldiol
3,3′-dihydroxyhexestrol
3′,4′,7-trihydroxy isoflavone
3-deoxyestradiol
3-deoxyestrone
3-hydroxyestra-1,3,5(10)-trien-16-one
3-methylestriol
3-phenylphenol
4-(benzyloxyl)phenol
4,4′-(1,2-ethanediyl)bisphenol
4,4′-dihydoxybenzophenone
4,4′-dihydroxy stibene
4,4′-sulfonyldiphenol
4′,6-dihydroxyflavone
4-chloro-2-methyl phenol
4-chloro-3-methylphenol
4-chloro-4′-biphenylol
4-cresol
4-dodecylphenol
4-ethyl-7-OH-3-(*p*-methoxyphenyl)di-hydro-1-benzopyran-2-one
4-ethylphenol
4-heptyloxyphenol
4-hydroxychalcone
4-hydroxybiphenyl
4′-hydroxychalcone
4-hydroxyestradiol
4-hydroxytamoxifen
4-*n*-octylphenol
4-phenethylphenol
4-*sec*-butylphenol
4-*tert*-amylphenol
4-*tert*-butylphenol
4-*tert*-octylphenol
6α-OH-estradiol
6-hydroxyflavanone
6-hydroxyflavone
7-hydroxyflavanone
α,α-dimethyl-β-ethyl allenolic acid
α-zearalanol
3α-androstanediol
3β-androstanediol
apigenin
aurin
baicalein
benzophenone, 2,4-hydroxy
benzyl 4-hydroxybenzoate
β-zearalanol
β-zearalenol
biochanin A
bis(4-hydroxyphenyl)methane
bisphenol A
bisphenol B
chalcone
clomiphene
coumestrol
daidzein
dienestrol
diethylstilbestrol
diethylstilbestrol dimethyl ether
diethylstilbestrol monomethyl ether
dihydrotestosterone
dihydroxymethoxychlor olefin
dimethylstibestrol
diphenolic acid
doisynoestrol
droloxifene
equol
estradiol
estriol
estrone
ethyl 4-hydroxybenzoate
ethynylestradiol
fisetin
3′-hydroxy flavanone
4′-hydroxy flavanone
3,6,4′-trihydroxy flavone
formononetin
genistein
heptyl p-hydroxybenzoate
hexestrol
HPTE
ICI 164384
ICI 182780
kaempferol
kepone
mestranol
methyl 4-hydroxybenzoate
m-ethylphenol
monohydroxymethoxychlor
monohydroxymethoxychlor olefin
monomethylether hexestrol
morin
moxestrol
myricetin
nafoxidine
naringenin
n-butyl 4-hydroxybenzoate
nonylphenol
nordihydroguaiaretic acid
norethynodrel
n-propyl 4-hydroxybenzoate
o,*p*′-DDT
phenol red
phenol, P−(α,β-diethyl-*p*-methylphenethyl)-,mes
p-cumyl phenol
phenolphthalein
phenolphthalin
phloretin
prunetin
rutin
tamoxifen
toremifene
triphenylethylene
zearalanone
zearalenol

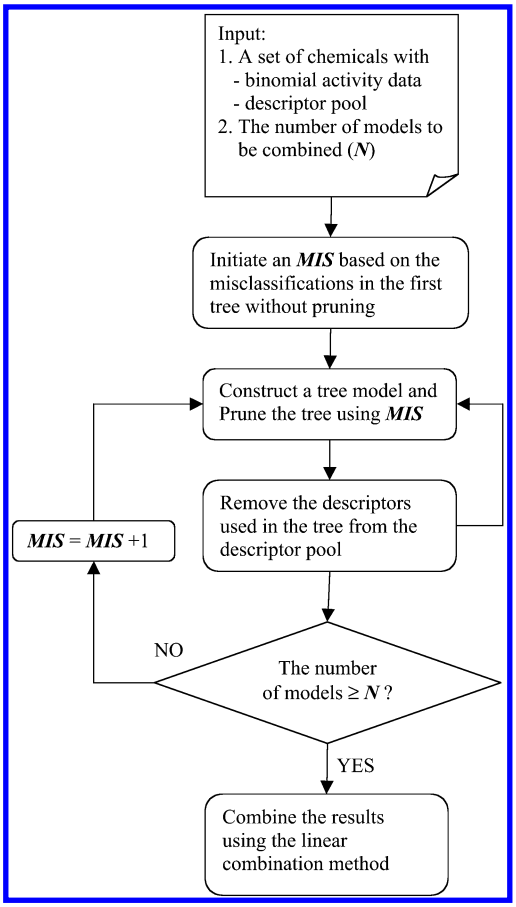

Figure 2. Flowchart of the Decision Forest algorithm. The parameter *MIS* determines the number of misclassified chemicals allowed in pruning.

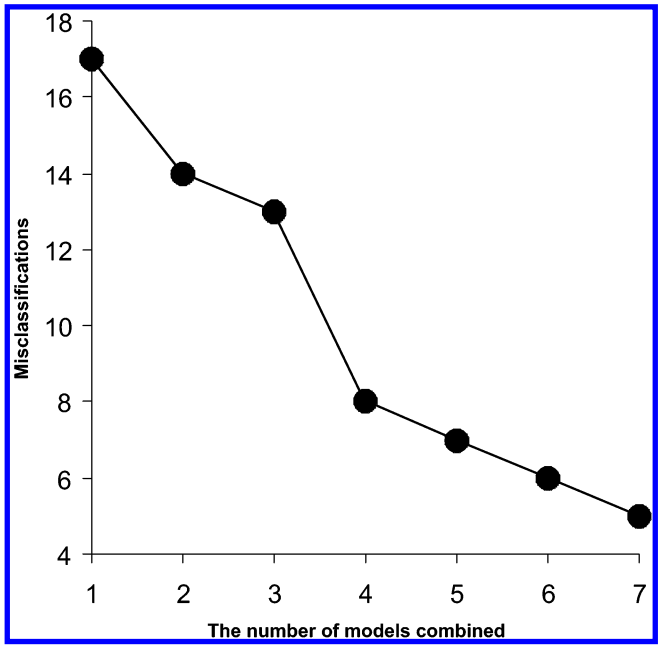

Figure 3. Relationship of misclassifications with the number of trees combined in Decision Forest.

Decision Forest can be understood as a pooling result of SAR models that predict activity within their unique chemistry spaces. Since each SAR model is developed using a unique set of descriptors, the difference in their prediction is maximized. Thus, it is safe to assume that combining multiple valid SAR models that use unique sets of descriptors into a single decision function should provide better estimation of activity than that separately predicted by the individual models.

Table 2. Results of Seven Individual Trees and Their Combination Performance

			misclassifications in	
tree ID	no. of descriptors used	no. of leafs	each tree	combination
1	10	13	17	17
2	10	13	19	14
3	12	15	17	13
4	12	14	17	8
5	15	18	19	7
6	16	19	20	6
7	13	17	18	5

Table 3. Comparison of Model Performance between Decision Tree and Decision Forest

		decision tree prediction[a]		decision forest prediction[a]	
		A	I	A	I
expt results	A = 131	122	9	127	4
	I = 101	8	93	1	100

[a] A = active; I = inactive.

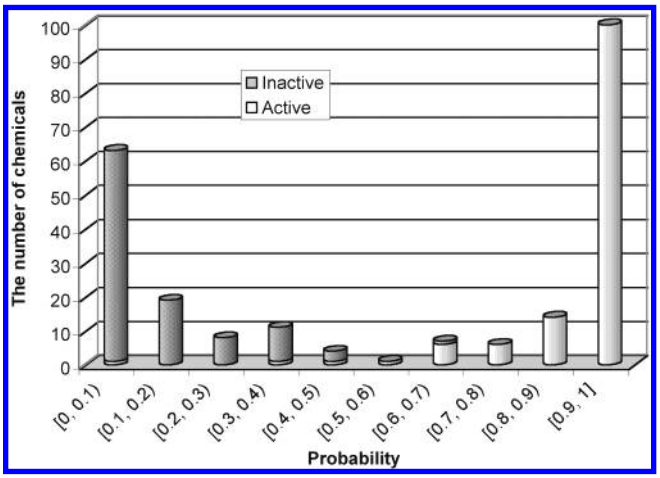

Figure 4. Distribution of active/inactive chemicals across the probability bins in Decision Forest. The probability of each chemical was the mean value calculated over all individual trees in Decision Forest. A chemical with probability larger than 0.5 was designated as active while less than 0.5 was inactive.

A number of commercial software packages, including CODESSA (Semichem, Shawnee, KS), Cerius2 (Accelrys Inc., San Diego, CA), and Molconn-Z (eduSoft, LC, Richmond, VA), enable a large volume of descriptors to be generated for SAR studies. Decision Forest takes advantage of this large volume of descriptors by aggregating the information of structural dependence on activity represented from each unique set of descriptors. Unlike the resampling techniques used in most combining forecast approaches, all training chemicals are included in each decision tree to be combined in the Decision Forest, thus maximizing the SAR information.

It is important to note that there is always a certain degree of noise associated with biological data and particularly the data generated from a HTS process. Thus, optimizing SAR models inherently risks over fitting the noise, a result most often observed using ANNs. Since the combination scheme

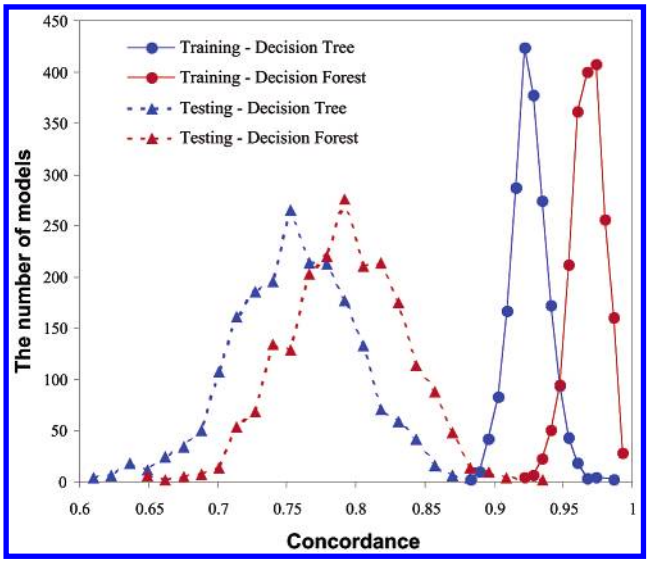

Figure 5. Comparison of the results between the Decision Tree and the Decision Forest model in a validation process. In this method, the data set was divided into two groups, 2/3 for training and 1/3 for testing. The process was repeated 2000 times. The red line is associated with the results from Decision Forest while the blue line is for Decision Tree. The quality of a model in both training (●) and predication (▲) was assessed using concordance that was calculated by dividing the misclassifications by the number of training chemicals in the training step and by the number of testing chemicals in prediction, respectively. The position of a dot (● or ▲) on the graph identifies the number of models with a certain value of concordance.

of Decision Forest is not a fitting process, some noise introduced by individual SAR models will be canceled when combining predictions. Moreover, using Decision Tree to construct Decision Forest offers additional benefits because the quality of a tree can be adjusted in the pruning process using the *MIS* parameter as a figure of merit for model quality. The *MIS* parameter is an indicator of noise, enabling the modeler a way to reduce over fitting of the noise.

Decision Forest can be used for priority setting in both drug discovery and regulatory applications. The objective of priority setting is to rank order from most important to least important a large number of chemicals for experimental evaluation. The purpose of priority setting in drug discovery is to identify a few lead chemicals but not necessarily all potential ones. In other words, relatively high false negatives are tolerable, but false positives need to be low. In the example we presented, chemicals predicted to be active with probability > 0.7 were shown to have 100% concordance with experimental data, thus demonstrating its use for lead selection.

In contrast, a good priority setting method for regulatory application should generate a small fraction of false negatives. False negatives constitute a crucial error, because they will receive a relatively lower priority for experimental evaluation. In the example we presented, chemicals predicted to be inactive with probability < 0.3 were shown to have 98.9% concordance with experimental data, thus demonstrating its use for regulatory application.

ACKNOWLEDGMENT

The research is funded under the Inter-Agency Agreement between the U.S. Environmental Protection Agency and the U.S. Food and Drug Administration's National Center for Toxicological Research. The authors also gratefully acknowledge the American Chemistry Council and the FDA's Office of Women's Health for partial financial support.

REFERENCES AND NOTES

(1) Rusinko, A., III; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Analysis of a large structure/biological activity data set using recursive partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017−1026.
(2) Hawkins, D. M.; Young, S. S.; Rusinko, A., III Analysis of large structure−activity data set using recursive partitioning. *Quant. Struct.-Act. Relat.* **1997**, *16*, 296−302.
(3) Morgan, J. N.; Sonquist, J. A. Problems in the analysis of survey data, and a proposal. *J. Am. Statist. Assoc.* **1963**, *58*, 415−434.
(4) Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. *Classification and regression trees*; Chapman and Hall: 1984.
(5) Breiman, L.; Friedman, J.; Olshen, R.; Stone, C.; Steinberg, D.; Colla, P. *Cart: Classification and regression trees*; 1995.
(6) Lim, T.-S.; Loh, W.-Y. *A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms*; Cohen, W. W., Ed.; Kluwer Academic Publishers: 1999; pp 1−27.
(7) Shi, L. M.; Tong, W.; Fang, H.; Perkins, R.; Wu, J.; Tu, M.; Blair, R.; Branham, W.; Walker, J.; Waller, C.; Sheehan, D. An integrated "4-Phase" approach for setting endocrine disruption screening priorities − Phase I and II predictions of estrogen receptor binding affinity. *SAR/QSAR Environ. Res.* **2002**, *13*, 69−88.
(8) Kass, G. V. An exploratory technique for investigating large quantities of categorical data. *Appl. Stat.* **1980**, *29*, 119−127.
(9) Quinlan, J. *C4.5: programs for machine learning*; Morgan Kauffman: 1993.
(10) Quinlan, J. R. Improved use of continuous attributes in C4.5. *J. Artif. Intel. Res.* **1996**, *4*, 77−90.
(11) Clark, L. A.; Pregibon, D. *Tree-based models*; Chambers & Hastie: 1997; Chapter 9, pp 413−430.
(12) Loh, W.-Y.; Vanichsetakul, N. Tree-structured classification via generalized discriminant analysis. *J. Am. Statist. Assoc.* **1988**, *83*, 715−728.
(13) Loh, W.-Y.; Shih, Y. S. Split selection methods for classification trees. *Statistica Sinica* **1997**, *7*, 815−840.
(14) Buntine, W.; Caruana, R. *Introduction to IND version 2.1 and recursive partitioning*; NASA Ames Research Center: 1992.
(15) Murthy, S. K.; Kasif, S.; Salzberg, S. A system for induction of oblique decision trees. *J. Artif. Intel. Res.* **1994**, *2*, 1−32.
(16) Brodley, C. E.; Utgoff, P. E. Multivariate decision trees. *Mach. Learn.* **1995**, *19*, 45−77.
(17) Muller, W.; Wysotzki, F. Automatic construction of decision trees for classification. *Ann. Oper. Res.* **1994**, *52*, 231−247.
(18) Muller, W.; Wysotzki, F. *The decision-tree algorithm CAL5 based on a statistical approach to its splitting algorithm*; Nakhaeizadeh, G., Taylor, C. C., Eds.; John Wiley & Sons: 1997; pp 45−65.
(19) Holte, R. C. *Very simple classification rules perform well on most commonly used datasets*; 1993; Vol. 11, pp 63−90.
(20) Blower, P.; Fligner, M.; Verducci, J.; Bjoraker, J. On combining recursive partitioning and simulated annealing to detect groups of biologically active compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 393−404.
(21) Izrailev, S.; Agrafiotis, D. A novel method for building regression tree models for QSAR based on artificial ant colony systems. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 176−180.
(22) Bates, J. M.; Granger, C. W. J. The combination of forecasts. *Oper. Res. Quart.* **1969**, *20*, 451−468.
(23) Opitz, D.; Shavlik, J. Actively searching for an effective neural-network ensemble. *Connect. Sci.* **1996**, *8*, 337−353.
(24) Krogh, A.; Vedelsby, J. *Neural network ensembles, cross validation and active learning*; Tesauro, G., Touretzky, D., Leen, T., Eds.; MIT Press: 1995; Vol. 7, pp 231−238.
(25) Maclin, R.; Shavlik, J. Combining the predictions of multiple classifiers: Using competitive learning to initialize neural networks. *Proc. 14th Int. Joint Conf. Intel.* **1995**, 524−530.
(26) Drucker, H.; Cortes, C. *Boosting decision trees*; MIT Press: 1996; Vol. 8, pp 479−485.
(27) Quinlan, J. Bagging, boosting and c4.5. *Proc. 13th Nat. Conf. Artif. Intel.* **1996**, 725−730.
(28) Bunn, D. W. *Expert use of forecasts: Bootstrapping and linear models*; Wright, G., Ayton, P., Eds.; Wiley: 1987; pp 229−241.
(29) Bunn, D. W. Combining forecasts. *Eur. J. Operat. Res.* **1988**, *33*, 223−229.

(30) Clemen, R. T. Combining forecasts: A review and annotated bibliography. *Int. J. Forecast.* **1989**, *5*, 559−583.
(31) Maclin, R.; Opitz, D. An empirical evaluation of Bagging and Boosting. *Proc. 14th Nat. Conf. Artif. Intel.* **1997**, 546−551.
(32) Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123−140.
(33) Freund, Y.; Schapire, R. Experiments with a new Boosting algorithm. *Proc. 13th Int. Conf. Mach. Learn.* **1996**, 148−156.
(34) Blair, R.; Fang, H.; Branham, W. S.; Hass, B.; Dial, S. L.; Moland, C. L.; Tong, W.; Shi, L.; Perkins, R.; Sheehan, D. M. Estrogen receptor relative binding affinities of 188 natural and xenochemicals: Structural diversity of ligands. *Toxicol. Sci.* **2000**, *54*, 138−153.
(35) Branham, W. S.; Dial, S. L.; Moland, C. L.; Hass, B.; Blair, R.; Fang, H.; Shi, L.; Tong, W.; Perkins, R.; Sheehan, D. M. Binding of phytoestrogens and mycoestrogens to the rat uterine estrogen receptor. *J. Nutrit.* **2002**, *132*, 658−664.
(36) Fang, H.; Tong, W.; Shi, L.; Blair, R.; Perkins, R.; Branham, W. S.; Dial, S. L.; Moland, C. L.; Sheehan, D. M. Structure activity relationship for a large diverse set of natural, synthetic and environmental chemicals. *Chem. Res. Toxicol.* **2001**, *14*, 280−294.
(37) Hong, H.; Tong, W.; Fang, H.; Shi, L. M.; Xie, Q.; Wu, J.; Perkins, R.; Walker, J.; Branham, W.; Sheehan, D. Prediction of estrogen receptor binding for 58,000 chemicals using an integrated system of a tree-based model with structural alerts. *Environ. Health Persp.* **2002**, *110*, 29−36.

CI020058S