Ensemble of Linear Models for Predicting Drug Properties

Tomasz Arodź,*,† David A. Yuen,‡ and Arkadiusz Z. Dudek§

Institute of Computer Science, AGH University of Science and Technology, 30-059 Kraków, Poland, Minnesota Supercomputing Institute, University of Minnesota, Minneapolis, Minnesota 55455, and University of Minnesota Medical School, Department of Medicine, Minneapolis, Minnesota 55455

Received September 5, 2005

We propose a new classification method for the prediction of drug properties, called random feature subset boosting for linear discriminant analysis (LDA). The main novelty of this method is the ability to overcome the problems with constructing ensembles of linear discriminant models based on generalized eigenvectors of covariance matrices. Such linear models are popular in building classification-based structure−activity relationships. The introduction of ensembles of LDA models allows for an analysis of more complex problems than by using single LDA, for example, those involving multiple mechanisms of action. Using four data sets, we show experimentally that the method is competitive with other recently studied chemoinformatic methods, including support vector machines and models based on decision trees. We present an easy scheme for interpreting the model despite its apparent sophistication. We also outline theoretical evidence as to why, contrary to the conventional AdaBoost ensemble algorithm, this method is able to increase the accuracy of LDA models.

1. INTRODUCTION

Computational methods in drug design have been studied extensively in past decades.[1,2] Major efforts are focused on the activity of therapeutic agents, as well as on other essential properties of pharmaceuticals. Such studies, aiming at predicting in a quantitative or qualitative manner the activity or property of a compound from its structure, are collectively known as structure−activity relationship (SAR) studies.[3] In particular, much attention in SAR research was directed toward predicting the pharmacokinetic properties of compounds, including absorption, distribution, metabolism, and excretion, as well as toxicity, collectively referred to as ADME/Tox.[4,5] Adverse drug effects are also studied, such as, for example, drug resistance.[6]

The SAR problems pose some specific requirements for the computational methods. The biochemical mechanisms of the studied effect are often unknown or diverse. Such multiple mechanisms of action can result in a highly nonlinear relationship between descriptors of the compound structure and its activity. The descriptors themselves pose another challenge. The large number of available types of descriptors makes it hard to manually choose only those meaningful in the context of a given activity. On the other hand, using too many descriptors results in chance correlations and in data sets with a large number of interdescriptor correlations. Next, the available information on the activity of compounds is usually small and may be highly unbalanced, with much fewer numbers of active than nonactive compounds. Moreover, as new compounds are created by combinatorial chemistry[7] on the basis of previous analyses, they may not follow the same distribution as the compound data sets previously available. Finally, the models resulting from SAR analysis should be, at the same time, accurate and easily interpretable. The SAR model should allow for reliable prediction of the activity of compounds during screening. However, it should also give insight into the properties of compounds that exhibit high activities, allowing for guiding the further exploration of chemical space. These two goals often contradict each other.

Various computational methods have been applied to the SAR analysis. A number of simple statistical methods are in widespread use, including linear models such as multiple linear regression,[8] linear discriminant analysis,[9,10] or partial least squares.[11] These models offer good interpretability, yet their accuracy is not always optimal, in particular for larger, more diverse data sets. More complex models are also in use, including artificial neural networks,[10] support vector machines (SVM),[12] or decision trees.[13] Recently, a group of ensemble methods gained attention from the chemoinformatic community. These models are multiple classifiers usually composed of decision trees, as in, for example, random forests[13,14] and stochastic gradient boosting of decision trees[15] methods. However, other ensemble types have also been tried in SAR, including SVM ensembles[16] and artificial neural networks with boosting[17] and bagging.[18]

Boosting methods, such as AdaBoost,[19] have been one of the most successful ensemble methods up to date, both empirically and in terms of theoretical foundations. However, they have only recently been introduced[15,17] to SAR studies and have been used with relatively complex base models, namely, decision trees[15] or artificial neural networks.[17] We propose here to use a simple linear LDA base model in a boosting framework. Previous analyses argued that the boosting of LDA does not lead to higher accuracy than a single LDA model.[20,21] Here, we introduce a new approach,

* Corresponding author phone: +48 12 617 3497; fax: +48 12 633 9406; e-mail: arodz@agh.edu.pl.
† AGH University of Science and Technology.
‡ Minnesota Supercomputing Institute, University of Minnesota.
§ University of Minnesota Medical School, Department of Medicine.

10.1021/ci050375+ CCC: $33.50
Published on Web 12/18/2005

random feature subset boosting (RFSBoost), which is able to overcome the problems with the boosting of LDA.

The new method is evaluated for four data sets related to drug absorption, resistance, and toxicity. These focus on predicting the human intestinal absorption of drugs and giving a prognosis whether a compound is a P-glycoprotein substrate or not and, thus, is susceptible to efflux from the cell. We have also studied the prediction of whether a compound induces Torsade de Pointes cardiac arrhythmia, an adverse side effect of some drugs, and the prediction of the multidrug resistance reversal activity of chemical agents.

The rest of the paper is arranged as follows. In Section 2, the RFSBoost of LDA is introduced. Moreover, the data sets used in this study are described, along with the computation procedure of evaluation of the method. In Section 3, we present and analyze the results and compare them with other SAR models used in the literature. Finally, Section 4 gives the conclusions.

2. METHODS

The ensemble and boosting schemes in machine learning have been extensively described in the literature, for example, by Freund and Schapire[22] and Meir and Rätsch.[23] Thus, we give only a brief description of boosting, focusing on the novel aspects of our method, its empirical and theoretical properties, and its interpretability. We discuss also the four SAR data sets analyzed.

2.1. Random Feature Subset Boosting for LDA. Similar to other ensemble methods, boosting[19] involves combining multiple models to obtain a more reliable model. Contrary to other ensembles, such as the random subspace method[24] or bagging,[25] the base models may be erroneous to a large extent. The high accuracy of the ensemble model is achieved by constructing new ensemble members to correct the errors of previously trained ensemble members. During the course of the algorithm, the weights of the examples from the training set are altered, shifting the attention of the next base classifiers to the previously misclassified examples. Each consecutive base model is trained to minimize the weighted error on the training set. Thus, a base classifier is forced to classify correctly the examples with large weights, even at the cost of making mistakes for examples with small weights.

However, for such a scheme to be successful, the base classifier must exhibit some instability with respect to changes in the weights. If, despite these changes, the base models are similar, the increase in accuracy is low in comparison to a single base model. The diversity of base models forming the ensemble is an important factor influencing the accuracy of the ensemble.[26]

The LDA used widely in SAR studies has been tested in boosting[20] and other ensemble types.[21] However, the results were discouraging. The ensemble of LDA models often gives results similar or worse than a single LDA model.[21] This is because models trained in consecutive rounds of boosting are very similar.[20] Therefore, to successfully apply the LDA as a base classifier, a method forcing the base models to be more diverse has to be used. Here, diversity is achieved by introducing the concept of random feature subsets to boosting. Specifically, in each turn, the new base classifier is trained using a different, randomly chosen subset of descriptors.

The algorithm for the RFSBoost for LDA method is outlined herein as Algorithm 1. First, the initial weights D_1 of all the samples in the training set are initialized uniformly for each class (line 1). Next, the algorithm operates sequentially in T rounds (line 2), in each, training a new base model. In every round, a group *FeatSset* of S features is chosen randomly from all features X (line 3). In the training of the LDA, the training set, with only the selected features, is used (line 4), finding the linear decision boundary with minimal weighted training error. The weighted version of the LDA method is used to take account of the weights D_t (line 5).

Once the new linear decision is obtained by LDA, the weighted classification error is evaluated (line 6) and the weight of the trained LDA model in the whole ensemble is calculated (lines 9−10). Next, the new weights of the examples, to be used in the next round, are derived (lines 11−12). The weights of the correctly classified examples are lowered, while those of misclassified examples remain unchanged. Finally, at the end of each round, the weights are normalized so that the sum of the weights for each of the classes is equal (line 13). This modification of the boosting scheme is introduced specifically for SAR analysis, where unbalanced class representations are not uncommon.

Algorithm 1 *The RFSBoost for LDA*

INPUT:
Training set Tr *of size* m*:*
$Tr = \{(\vec{x}_1, c_1), \ldots, (\vec{x}_m, c_m)\} \subset X \times \{-, +\}$
T *- maximal number of training rounds*
S *- number of features to be used*

RandomFeatureSubsetBoost − LDA (Tr, T, S)
1 $\forall_{1 \le i \le m} D_1(i) = \frac{1}{2} \frac{1}{\sum_{j:c_j = c_i} D_t(j)}$
2 **for** $t = 1$ **to** T
3 $FeatSset$ = **choose randomly** S **distinct features from** X
4 $TrFtSset$ = **restrict** Tr **to selected features** $FeatSset$
5 $h_t = \mathrm{LDA}(TrFtSset, D_t)$
6 $\varepsilon_t = \sum_{i:h_t(\vec{x}_i) \neq c_i} D_t(i)$
7 **if** $\varepsilon_t = 0$ **or** $\varepsilon_t \ge 0.5$
8 **exit loop**
9 $\beta_t = \frac{\varepsilon_t}{1 - \varepsilon_t}$
10 $\alpha_t = \log \frac{1}{\beta_t}$
11 $\forall_{i:h_t(\vec{x}_i) = c_i} D_{t+1}(i) = \beta_t D_t(i)$
12 $\forall_{i:h_t(\vec{x}_i) \neq c_i} D_{t+1}(i) = D_t(i)$
13 $\forall_{1 \le i \le m} D_t(i) = \frac{1}{2} \frac{D_t(i)}{\sum_{j:c_j = c_i} D_t(j)}$
14 **return** $h_{\mathrm{fin}}(\vec{x}) = \frac{\sum_{i=1}^{T} \alpha_i h_i(\vec{x})}{\sum_{i=1}^{T} \alpha_i}$

Our previous analysis of the RFSBoost−LDA method in a general classification setup[27] has given insight into the reasons the LDA works significantly better in the RFSBoost scheme than in conventional boosting. First, the diversity of the constructed RFSBoost ensemble, measured with the Q statistics[28] or with the variance of the decision boundary coefficients,[27] tends to be higher than that of AdaBoost. High diversity has been argued to influence the accuracy of the ensemble classifiers.[29]

Another justification stems from the margin-based analysis of boosting.[30] The margin of the ensemble on an example describes the distance from the decision boundary to that

example. The sign of the margin shows whether the example has been classified correctly, while the magnitude of the margin represents the confidence in the decision for that example. One of the reasons for the good performance of AdaBoost in general is the maximization of the margins. This leads to an increase in the confidence of the classification and a reduction of the number of errors. Compared to AdaBoost, the minimal margin of the RFSBoost ensemble of LDA models is, in most cases, larger, indicating better classification capabilities.

Finally, experimental results gain support from the statistical learning theory. The bound on the generalization error of boosting is dependent[19] on the Vapnik−Chervonenkis (VC) dimension[31] ϑ of the base classifier. For a linear classifier in the feature space defined by f descriptors, this dimension is $\vartheta_{\mathrm{LDA}} = f + 1$. In RFSBoost−LDA, the number of features used in LDA training is significantly reduced compared to AdaBoost−LDA. Thus, the VC dimension of each weak classifier is reduced and the bound on the generalization error of the whole ensemble is lower, leading to better performance during the classification of previously unseen samples.

RFSBoost exhibits other interesting properties not directly related to its accuracy. First, the linear methods such as LDA are not directly applicable in cases where the within-class covariance matrix is ill-conditioned. This includes, for example, data sets with a large descriptor-to-examples ratio. By using only a small fraction of the available descriptors in each round, RFSBoost allows the creation of ensembles of linear models in such cases. Another benefit of using subsets of descriptors is the significant reduction in the computational complexity of a linear base model, which compensates for the multiplicity of base models in an ensemble. Moreover, relying on subsets of features makes the method independent, to a large extent, of the number of descriptors provided by the investigator.

2.2. Interpreting the Ensemble of Linear Models. While achieving high accuracy is an essential goal of a SAR model, its interpretability is no less important. In particular, linear models offer good insight into the relation between compound descriptors and the analyzed drug activity or property. In more sophisticated models, based on machine learning, obtaining information on the importance of the descriptors is not as straightforward. In our method, each of the linear base models h_t contributing to the ensemble can be analyzed individually in a convenient way, by inspecting the coefficients h_t^f corresponding to the descriptor f. Since, in each round, only a group of descriptors is chosen randomly to participate in creating the base models, only the coefficients for those descriptors may be nonzero. In addition to the coefficients, the weight α_t of the tth base model is available for inspection.

While the above method allows for insight into small ensemble models, it is nonetheless prohibitively laborious for typical ensembles consisting of 100 or more base models. Therefore, we propose a scheme for a holistic analysis of the ensemble. The method takes into account the random process of choosing the subset of features in each round t. It also uses the information on the features' coefficients of each linear base model.

Table 1. Data Sets Used in Tests[a]

data set	test sets	training sets	class +	class −
HIA	157	39	131	65
P-gp	161	40	116	85
TdP	289	72	85	276
MDRR	396	132	298	230

[a] The sizes of the test and training sets during cross-validation are presented, along with a distribution of compounds between two classes to be predicted. See the text for details on the data sets and the meaning of "+" and "−" classes.

Let $p_t(f)$: $\mathbb{N} \rightarrow \{0, 1\}$ be a function taking 1 if a descriptor f is selected by the random process in the base model h_t and 0 otherwise. Then, $c(f) = \sum_{t=1}^{T} p_t(f)$ defines the number of rounds from the total T rounds, in which the descriptor f is used to build the base model. The importance of the descriptor f in a base model h_t is given by the magnitude of its coefficient, $||h_t^f||$. Therefore, the estimate of the importance of a descriptor f in an ensemble of T base models is

$$I(f) = \frac{1}{c(f)} \sum_{t=1}^{T} \alpha_t ||h_t^f|| \qquad (1)$$

The importance factor I takes into account the importance of descriptors within base models and of the base models within the ensemble. By utilizing $c(f)$ factors, it also compensates for the random nature of the choice of descriptors in creating each base model.

2.3. Data Sets. We have evaluated the RFSBoost for LDA ensemble model using four prediction problems encountered in drug design. The number of compounds in each data set and their distribution into classes of compounds are summarized in Table 1.

Human Intestinal Absorption (HIA). The HIA prediction focuses on the absorption of orally administered drugs into blood. A high absorption rate is essential for bioavailability of the drug at its target site.[32] For the purpose of prediction, the compounds with absorption rates above 70% were classified as absorbable (HIA +) and the rest as nonabsorbable (HIA −). The data set has been obtained from Xue et al.[33] and contains 131 absorbable and 65 nonabsorbable compounds.

P-glycoprotein (P-gp). The second data set involves the prediction of P-pg substrates from nonsubstrates. P-gp is a membrane transporter capable of transporting out of the cell a wide and diverse range of chemical compounds, including many therapeutic agents.[34] Thus, it is important in drug design to evaluate, in an early stage, whether a drug is not a P-gp substrate susceptible to P-gp-mediated drug efflux from the cell, limiting its activity. The data set under investigation has been collected from various literature sources by Xue et al.[33] and includes 116 P-gp substrates (P-gp +) and 85 nonsubstrates (P-gp −).

Torsade de Pointes (TdP). Torsade de Pointes is a potentially fatal polymorphic ventricular tachycardia.[35] It is linked to genetic causes leading to long QT intervals on the electrocardiogram and has been studied both physiologically[36] and computationally.[37] It may also be induced as an adverse effect of drugs that cause QT prolongation. This effect is present in different categories of therapeutic agents, for example, antihistamines, antidepressants, or macrolide

antibiotics.[38] The TdP data set has been collected by Xue et al.[33] It includes 85 TdP-inducing agents (TdP +) and 276 noninducing compounds (TdP −).

Multidrug Resistance Reversal (MDRR). The MDDR activity prediction is linked to the P-gp transporter described above. In response to a single cytotoxic drug, the MDR1 gene encoding P-gp may become upregulated[39] and, thus, lead to resistance to a range of structurally and functionally unrelated drugs by means of their efflux from the cell. This leads to major problems, for example, in cancer chemotherapy. One of the promising approaches to tackle this problem is the use of MDDR agents. The set of 528 compounds originally studied by Klopman et al.[6] and Bakken and Jurs[40] has been obtained from Svetnik et al.[13] It includes 298 active (MDRR +) compounds, with a ratio of reversing the leukaemia cells' resistance to adriamycin above 4.2, and 230 inactive (MDRR −) compounds, with this ratio below 2.0.

Molecular Descriptors. For HIA, P-gp, and TdP data sets, a group of 159 molecular descriptors is used from Xue et al.[33] It includes 18 simple molecular properties, 28 molecular connectivity and shape descriptors, 84 electrotopological state descriptors, 18 quantum chemical properties, and 16 geometrical properties. In the case of the MDRR data set, 342 DRAGON descriptors, previously used by Svetnik et al.,[13] are employed.

2.4. Computation Procedure. To evaluate the accuracy of the RFSBoost−LDA method, we have averaged the results over 10 runs of the algorithm on different random partitions of the data sets into training and test sets. The exact numbers of compounds in the training and test sets are specified in Table 1. On the basis of the results on the test sets, we have calculated four values quantifying the model's reliability. These are

- Accuracy (Acc): the percentage of correctly predicted samples over the total number of samples in the test set.
- Sensitivity (S_e): the percentage of correctly predicted samples from class "+" over the total number of samples from class "+" in the test set.
- Specificity (S_p): the percentage of correctly predicted samples from class "−" over the total number of samples from class "−" in the test set.
- Balanced accuracy (Acc_b), as used in the SAR study by Weston et al.[41]: the average of S_e and S_p. During the evaluation, two sizes of the ensembles were used, consisting of $T = 100$ or 200 base models. For a given T, the RFSBoost behavior is controlled by a free parameter S, specifying the number of features to be randomly chosen in each round. For all four data sets, we have tested configurations with $S = 2$, 4, 6, 8, and 10 descriptors. To select the optimal value of T for a given data set, we have used $T = 100$ if the models for this T achieved training accuracies close to 100% or $T = 200$ otherwise. Thus, we give preference to the less complex ensembles, provided they are sufficiently accurate. From all configurations with different number of features, S, we have selected the one yielding the highest accuracy on the training set as the final result of the tests. In the case of several configurations achieving equal training accuracy, the one yielding a larger margin during training was chosen.

Matlab 7.0 R14 was used to solve the generalized eigenproblem of the LDA method. The first generalized eigenvector of within- and between-class covariance matrices leads to the optimal linear transformation that separates the classes while reducing the within-class variances. The rest of the RFSBoost algorithm was developed in-house using C++ and Matlab 7.0 R14.

Table 2. Averaged Results of the Tests of the Random Feature Subset Boosting of LDA[a]

data set	Acc	S_e	S_p	Acc_b
HIA	80.8	86.9	68.5	77.7
P-gp	76.3	83.0	67.1	75.1
TdP	74.2	72.9	74.5	73.7
MDRR	81.8	89.2	72.1	80.7

[a] Ensembles of 100 base models for HIA, P-gp, and TdP and 200 for MDRR. See the text for details on the data sets. Acc = accuracy, S_e = sensitivity, S_p = specificity, Acc_b = balanced accuracy. All results are in percents.

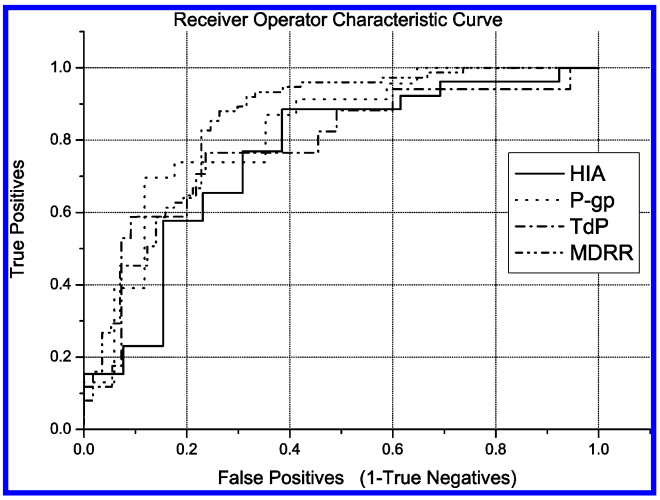

Figure 1. Receiver operator characteristics curve for all data sets, representing the tradeoff between true positives (sensitivity) and false positives (equal to 1 − specificity) on the test sets.

3. RESULTS AND DISCUSSION

The results of the RFSBoost evaluation are gathered in Table 2. In HIA, P-gp, and MDRR data sets, the accuracies for the classes were uneven, with a higher accuracy for the class with more examples, as can be observed by comparing specificity and sensitivity values.

For HIA and P-gp, the optimal value of S on the training set was 10 descriptors; for TdP, it was 4; and for MDRR, it was 8 descriptors. Models with those values of S were used in a further analysis and in comparisons. The ensembles for HIA, P-gp, and TdP consisted of 100 base models. The ensemble for MDRR, which is a data set with more compounds and features, was composed of 200 base models.

To illustrate the tradeoff between sensitivity and specificity on the test set, we have used the receiver operator characteristics (ROC) curve depicted in Figure 1, recently used in a chemoinformatic analysis by Müller et al.[42] The diagram is constructed using the model yielding a median accuracy from 10 models trained during evaluation for the chosen S.

We also show the results of the prediction using pharmacological distribution diagrams (PDDs) introduced by Galvez et al.[43] The diagrams for HIA and MDRR are presented in Figures 2 and 3, respectively. The diagrams were prepared using the combined results of the evaluation on the test sets.

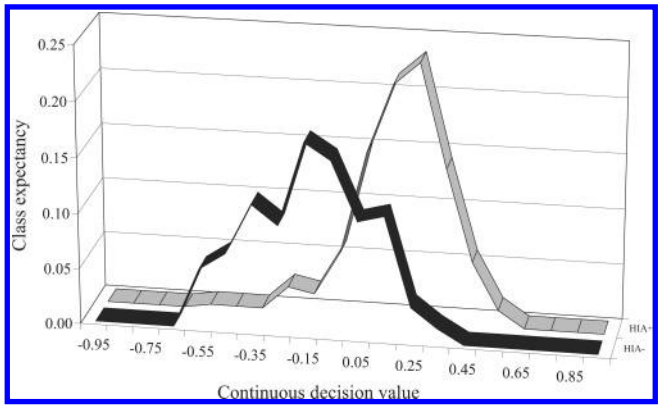

Figure 2. Pharmacological distribution diagram for the HIA data set. Active and inactive expectancy[43] as a function of the prethreshold continuous decision value of compounds from the test set.

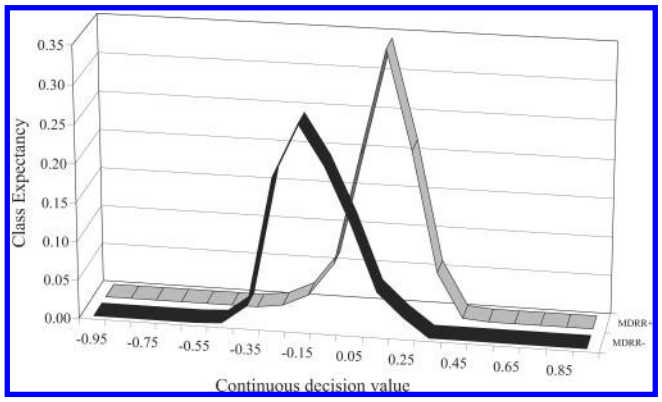

Figure 3. Pharmacological distribution diagram for the MDRR data set. Coordinates are as in Figure 2.

The ROC curve and the PDDs show the possibility of manipulating the decision threshold in RFSBoost to increase the accuracy within one class at the expense of the other class. Such an analysis is especially useful in the context of uneven costs of misclassification or an unbalanced nature of the data set, both situations being encountered in SAR analyses.

We have also analyzed the statistical quality of the created models. To this end, we have used the permutation test for classification problems,[44] by executing random scrambling of the dependent variable 250 times. For each permutation, the accuracy was averaged over 10 runs of evaluation, using the same partitionings into training and testing sets as in the evaluation above. Next, we created the cumulative distribution function of the averaged training error. For HIA and MDRR data sets, these distributions are depicted in Figure 4. Finally, we have calculated the p value, that is, the value of the cumulative distribution function corresponding to the averaged training error on the real, unscrambled data. For the inspected data sets, the p value is 0.028, 0.008, 0.048, and 0.004 for HIA, P-gp, TdP, and MDRR, respectively. As can be observed, the probabilities that the models are statistical artifacts are negligible. This suggests the models capture the actual structure−activity relationships.

3.1. Effects of Training Set Size. Having a small number of compounds with available assessed activity is a common situation in QSAR studies. In particular, data sets comprising analogues of a single hit are often composed of 100 or less compounds. While complex, nonlinear methods are certainly useful in large libraries, for small, more homogeneous data sets, linear models may suffice. To prove the usefulness of the nonlinear RFSBoost−LDA also in the latter case, we have evaluated its accuracy during the reduction of the training set size. On the basis of the HIA data set, we have constructed a series of prediction problems with a reduced number of training compounds, ranging from 30 to 70% of the whole data set. For each inspected training set size, we have generated 10 random subsets of each of the replicated training sets used in the evaluation described above.

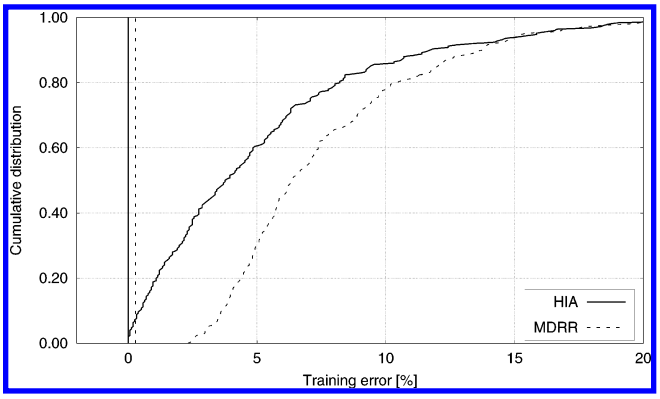

Figure 4. Results of the permutation tests for the HIA and MDRR data sets. Cumulative distribution functions of models achieving given training error while operating on scrambled data. Training errors for unscrambled data are marked with vertical lines.

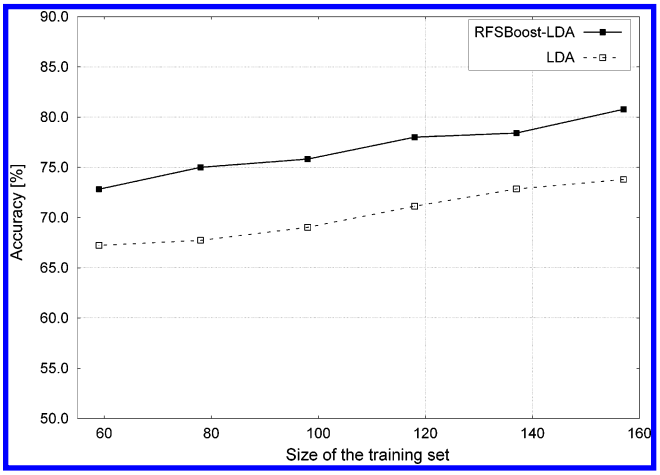

Figure 5. Comparison of the test set accuracies of RFSBoost and linear LDA models built using training sets of sizes reduced successively from 80% to 70%, 60%, 50%, 40%, and 30% of the whole HIA data set. The number of base models in the ensemble followed the reduction in training set size, with $T = 80$, 60, 40, 20, and 10, respectively.

We have preserved the same test sets and used the optimal parameters of the classifiers as specified above, with the exception of the number of rounds, T. To take account of the smaller data sets, we have executed lower numbers of rounds in boosting, ranging from $T = 10-80$. The results are depicted in Figure 5. The accuracy of RFSBoost−LDA is compared with that of the linear LDA model, operating in principal component space to prevent singularity due to the number of descriptors.

The results show that, while the accuracy of RFSBoost drops when training information is being limited, a similar accuracy decline can be observed for the linear method. The edge in prediction accuracy of the proposed method over simpler, linear models is, thus, not limited to larger data sets.

Table 3. Descriptors with Highest Importance Factor *I* (See eq 1) Across the 10 Executions of RFSBoost−LDA on the Same Training Set from the HIA Data Set[a]

N	descriptor	
10	VS	van der Waals surface area
10	μ	molecular dipole moment
10	nhyd	count of hydrogen atoms
10	nhev	count of heavy atoms
10	$^0\chi$	simple molecular connectivity for path order 0
10	$^2\chi$	simple molecular connectivity for path order 2
10	$^0\chi^v$	valence molecular connectivity for path order 0
10	$^1\kappa$	molecular shape κ index for one-boned fragments
10	$^1\kappa_\alpha$	κ_α index for one-boned fragments
10	$^2\kappa_\alpha$	κ_α index for two-boned fragments
9	ϵ_a	hydrogen bond donor acidity
9	A	electron affinity
9	$^1\chi$	simple molecular connectivity for path order 1
9	S(22)	atom-type H Estate sum for $>CH_2$
8	$^1\chi^v$	valence molecular connectivity for path order 1
8	S(38)	atom-type H Estate sum for :N:−

[a] *N* = number of executions for which the descriptor is among the 20 descriptors of highest importance.

3.2. Interpreting the Ensemble Model. To show the possibilities of interpreting the ensemble model, we have used the results obtained for the HIA data set. We have evaluated the importance *I*(*f*) of each descriptor *f* as described in eq 1. The training of the ensemble contains random elements. Thus, we have run the RFSBoost−LDA algorithm 10 times on the same training set. From each of the 10 trained models, we have selected 20 descriptors with the highest importance factors *I*.

Despite the reliance of training on a randomized procedure, within the 10 pools of 20 highest-ranking descriptors, 10 descriptors were present in the results of all 10 executions of training, 4 features were in 9 runs, and further 2 descriptors were in 8 pools. Some other descriptors, for example, the number of H-bond donors, were present in less than eight pools. This shows that the importance factor of the descriptors, in the form defined in eq 1, is independent, to a large extent, of the stochastic nature of training. The most important descriptors, along with information on the number of pools that they were present in, is presented in Table 3.

3.3. Comparison with Other Methods. The results obtained for the RFSBoost−LDA method were compared with other methods. To facilitate such a comparison, in our study, we have used the same descriptors as those in the referenced studies. The compounds used and the number of compounds in the training and test sets are also preserved. The results obtained by Xue et al.[33] for HIA, P-gp, and TdP using support vector machines are presented in Table 4, along with ones obtained by Svetnik et al.[15] for MDRR using decision trees, support vector machines, partial least squares, the stochastic gradient boosting of decision trees, and random forests.

For the HIA and P-gp data sets, our ensemble model has achieved an accuracy higher by 3.8% and 8%, respectively, than that of the SVM operating on all descriptors. It also yielded a balanced accuracy higher by 4.4% and 6.5%, respectively. While, in the case of TdP, the SVM yielded an accuracy higher by 7.8% than that of our model, it shows a significant lack of balance between sensitivity and specificity. When balanced accuracies Acc_b are compared, our ensemble model fares better by 1.1%. Moreover, the methodology of Xue et al. involved choosing the best model during parameter tuning using the error on the prediction set, while we have chosen the more conservative scenario of using the training error.

Table 4. Results of Other Studies Involving the Same Compounds and Descriptors and a Similar Computational Procedure as Used In Our Test[a]

model	Acc	S_e	S_p	Acc_b
		HIA		
SVM	77.0	83,4	63,2	73,3
		P-gp		
SVM	68.3	68.9	68.2	68.6
		TdP		
SVM	82.0	54.5	90.6	72.6
		MDRR		
DT	77.9			
SVM	81.5			
PLS	81.7			
SGB	82.6			
RF	83.1			

[a] Results after Xue et al.[33] for HIA, P-gp, and TdP and after Svetnik et al.[15] for MDRR. Acc = accuracy, S_e = sensitivity, S_p = specificity, Acc_b = balanced accuracy. All results are in percents.

The comparison of results for MDRR is restricted to accuracy, as Svetnik et al.[15] did not quote the sensitivity and specificity. RFSBoost−LDA is comparable in performance to other methods, with differences in accuracy ranging from 3.9% in favor of RFSBoost in the case of decision trees to 1.3% in favor of random forests.

These results show that the model we have proposed is among the best of classification models when evaluated on the same data set and using all descriptors. In the cases of HIA, P-gp, and TdP, Xue et al.[33] also tested a method consisting of recursive feature elimination (RFE) and SVM to reduce the number of descriptors. This method achieved an accuracy of 86.7% for HIA, 79.4% for P-gp, and 83.9% for TdP, which are higher than the results of our models. One should note that the RFSBoost for LDA method is capable of achieving results closer in accuracy to those of RFE−SVM when significantly larger ensembles are used. To show this, we evaluated our method for ensembles of $T = 2000$ base models and followed Xue et al.[33] in showing the configuration yielding the best results on the test set. We have achieved an accuracy of 82.1% for HIA, 80.8% for P-gp, and 79.2% for TdP.

The dedicated descriptor selection techniques, such as RFE, treat the reduction of irrelevant features as one of their main objectives. On the other hand, the random feature subsets technique serves a different purpose in our model. It is used for destabilizing the LDA base models. The design of a technique that allows for discarding irrelevant descriptors and, at the same time, destabilizes the LDA base models, thus, seems an interesting future research goal.

For HIA, P-gp, and MDRR prediction, other studies, not directly comparable with ours, have been performed. These studies use different sets of compounds and descriptors, and thus, the accuracy values may only give some general orientation on their relation to our method. For HIA prediction, the accuracy of a probabilistic neural network[45] was 80%. Much smaller training and testing sets were used,

with a lower threshold between absorbable and nonabsorbable compounds. In the case of P-gp substrate prediction, a multipharmacophore model[46] on a slightly smaller data set achieved an accuracy of 63%. The same set of compounds, with minor modifications, has been studied also by Svetnik et al.[13,15] Using a different set of descriptors, namely, the binarized atom-pair descriptors, an accuracy reaching 75.5% for stochastic gradient boosting and 80.4% for decision forests has been achieved. Finally, the MDRR problem was studied by Bakken and Jurs[40] on the same set of compounds and interclass separation thresholds, with a similar but not identical set of descriptors. Their method, utilizing a genetic algorithm for descriptor selection and a LDA method, achieved 83.1% accuracy. In general, all these studies show lower or similar accuracy to ours.

4. CONCLUSIONS

In this work we have proposed the novel random feature subset boosting of LDA. The method allows for overcoming the problem of low accuracy in ensembles composed of linear discriminant analysis as base models. This result is achieved by destabilizing, in each turn, the LDA base model by using only a small, randomly chosen subset of features. Thus, our approach unites two concepts, random subspacing and boosting, which on their own fail to create accurate LDA ensembles. The key to success in our method lies in the increase of diversity obtained by introducing random feature subsets, while keeping the strong drive for accuracy present in conventional boosting.

We have shown experimentally that our method is competitive in terms of accuracy with other classification methods recently proposed in SAR studies, such as support vector machines and random forests. The method is computationally efficient, owing to the multifold reduction in the number of features used in creating each linear method. Moreover, accurate results are obtained using ensembles of moderate size, composed of 100−200 base models, whereas, for example, successful applications of random forests employed 500 base trees.[13] In another study, the boosting of decision trees utilized optimization of the number of base models up to 500 or 1000 trees, depending on the size of the data set.[15]

Compared to linear methods,[47] RFSBoost remains competitive even for small data sets. Tests with reduced training set sizes indicated lower prediction errors than LDA, even when few compounds were available to the methods. The strength of linear models in such cases stems significantly from their evident interpretability. Interpreting complex, nonlinear models is usually not as straightforward. Yet, for small data sets, where minute ensembles suffice, RFSBoost allows for insight into the model by analyzing its linear constituents. For larger ensembles, we have shown a computationally efficient method for assessing the importance of each descriptor in the combined model.

Finally, we have noted that the use of a dedicated descriptor elimination method allows the SVM to elevate accuracy beyond that of our model on some data sets. Thus, we have proposed a future research goal of introducing a descriptor selection method to be used in conjunction with random feature subsets.

ACKNOWLEDGMENT

The authors thank Prof. Witold Dzwinel for inspiring discussions and for insightful comments during the preparation of the manuscript. The research is supported by CMG of National Science Foundation and by the Polish Ministry of Science and Information Society Technologies Grant 3 T11F 019 29.

Supporting Information Available: The pharmacological distribution diagrams for P-gp and TdP data sets. Details for the permutation tests and resulting diagrams for P-gp and TdP data sets. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Gershell, L. J.; Atkins, J. H. A brief history of novel drug discovery technologies. *Nat. Rev. Drug Discovery* **2003**, *2*, 321−327.
(2) Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303*, 1813−1818.
(3) Debnath, A. K. Quantitative Structure−Activity Relationship (QSAR) paradigm − Hansch era to new millennium. *Mini Rev. Med. Chem.* **2001**, *1*, 187−195.
(4) Ekins, S.; Rose, J. In silico ADME/Tox: the state of the art. *J. Mol. Graphics Modell.* **2002**, *20*, 305−309.
(5) van de Waterbeemd, H.; Gifford, E. ADMET in silico modelling: towards prediction paradise? *Nat. Rev. Drug Discovery* **2003**, *2*, 192−204.
(6) Klopman, G.; Shi, L. M.; Ramu, A. Quantitative structure−activity relationship of multidrug resistance reversal agents. *Mol. Pharmacol.* **1997**, *52*, 323−334.
(7) Bleicher, K. H.; Bohm, H.-J.; Muller, K.; Alanine, A. I. Hit and lead generation: Beyond high-throughput screening. *Nat. Rev. Drug Discovery* **2003**, *2*, 369−378.
(8) Farkas, O.; Heberger, K. Comparison of ridge regression, partial least-squares, pairwise correlation, forward- and best subset selection methods for prediction of retention indices for aliphatic alcohols. *J. Chem. Inf. Model.* **2005**, *45*, 339−346.
(9) Galvez, J.; Garcia-Domenech, R.; de Julian-Ortiz, J.; Soler, R. Topological approach to drug design. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 272−284.
(10) Murcia-Soler, M.; Perez-Gimenez, F.; Garcia-March, F. J.; Salabert-Salvador, M. T.; Diaz-Villanueva, W.; Castro-Bleda, M. J.; Villanueva-Pareja, A. Artificial neural networks and linear discriminant analysis: A valuable combination in the selection of new antibacterial compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1031−1041.
(11) Wold, S.; Sjostrom, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.
(12) Burbidge, R.; Trotter, M.; Buxton, B. F.; Holden, S. B. Drug design by machine learning: Support Vector Machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26*, 5−14.
(13) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, C.; Sheridan, R. P.; Feuston, B. P. Using general regression and probabilistic neural networks to predict human intestinal absorption with topological descriptors derived from two-dimensional chemical structures. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947−1958.
(14) Guha, R.; Jurs, P. C. Development of linear, ensemble, and nonlinear models for the prediction and interpretation of the biological activity of a set of PDGFR inhibitors. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2179−2189.
(15) Svetnik, V.; Wang, T.; Tong, C.; Liaw, A.; Sheridan, R. P.; Song, Q. Boosting: An ensemble learning tool for compound classification and QSAR modeling. *J. Chem. Inf. Model.* **2005**, *45*, 786−799.
(16) Merkwirth, C.; Mauser, H.; Schulz-Gasch, T.; Roche, O.; Stahl, M.; Lengauer, T. Ensemble methods for classification in cheminformatics. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2179−2189.
(17) He, P.; Xu, C.-J.; Liang, Y.-Z.; Fang, K.-T. Improving the classification accuracy in chemistry via boosting technique. *Chemom. Intell. Lab. Syst.* **2004**, *70*, 39−46.
(18) Agrafiotis, D. K.; Cedeno, W.; Lobanov, V. S. On the use of neural network ensembles in QSAR and QSPR. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 903−911.
(19) Freund, Y.; Schapire, R. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119−139.
(20) Skurichina, M.; Duin, R. P. W. Boosting in linear discriminant analysis. *Lect. Notes Comput. Sci.* **2000**, *1857*, 190−199.

(21) Skurichina, M.; Duin, R. P. W. Bagging, boosting and the random subspace method for linear classifiers. *Pattern Anal. Appl.* **2002**, *5*, 121−135.
(22) Freund, Y.; Schapire, R. A short introduction to boosting. *J. Jpn. Soc. Artif. Intell.* **1999**, *14*, 771−780.
(23) Meir, R.; Rätsch, G. An introduction to boosting and leveraging. *Lect. Notes Comput. Sci.* **2003**, *2600*, 118−183.
(24) Ho, T. K. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 832−844.
(25) Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123−140.
(26) Kuncheva, L. That elusive diversity in classifier ensembles. *Lect. Notes Comput. Sci.* **2003**, *2652*, 1126−1138.
(27) Arodź, T. Boosting the Fisher Linear Discriminant with random feature subsets. In *Computer Recognition Systems*, *Advances in Soft Computing*; Springer: Berlin, Heidelberg, 2005.
(28) Kuncheva, L. I.; Whitaker, C. J. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach. Learn.* **2003**, *51*, 181−207.
(29) Brown, G.; Wyatt, J.; Harris, R.; Yao, X. Diversity creation methods: A survey and categorisation. *Inf. Fusion* **2005**, *6*, 5−20.
(30) Schapire, R. E.; Freund, Y.; Bartlett, P.; Lee, W. S. Boosting the margin: a new explanation for the effectiveness of voting methods. *Ann. Stat.* **1998**, *26*, 1651−1686.
(31) Vapnik, V. *Estimation of Dependences Based on Empirical Data: Springer Series in Statistics*; Springer: New York, 1982.
(32) Wessel, M. D.; Jurs, P. C.; Tolan, J. W.; Muskal, S. M. Prediction of human intestinal absorption of drug compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 726−735.
(33) Xue, Y.; Li, Z. R.; Yap, C. W.; Sun, L. Z.; Chen, X.; Chen, Y. Z. Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1630−1638.
(34) Seelig, A. A general pattern for substrate recognition by P-glycoprotein. *Eur. J. Biochem.* **1998**, *251*, 252−261.
(35) Layton, D.; Key, C.; Shakir, S. A. Prolongation of the QT interval and cardiac arrhythmias associated with cisapride: limitations of the pharmacoepidemiological studies conducted and proposals for the future. *Pharmacoepidemiol. Drug Saf.* **2003**, *12*, 31−40.
(36) Keating, M. T. The long QT syndrome: A review of recent molecular genetic and physiologic discoveries. *Medicine* **1996**, *75*, 1−5.
(37) Saucerman, J. J.; Healy, S. N.; Belik, M. E.; Puglisi, J. L.; McCulloch, A. D. Proarrhythmic consequences of a KCNQ1 AKAP-binding domain mutation. Computational models of whole cells and heterogeneous tissue. *Circ. Res.* **2004**, *95*, 1216−1224.
(38) Ponti, F. D.; Poluzzi, E.; Cavalli, A.; Recanatini, M.; Montanaro, N. Safety of nonantiarrhythmic drugs that prolong the QT interval or induce torsade de pointes: an overview. *Drug Safety* **2002**, *25*, 263−286.
(39) Gottesman, M.; Pastan, I. Biochemistry of multidrug resistance mediated by the multidrug transporter. *Annu. Rev. Biochem.* **1993**, *62*, 385−427.
(40) Bakken, G. A.; Jurs, P. C. Classification of multidrug-resistance reversal agents using structure-based descriptors and linear discriminant analysis *J. Med. Chem.* **2000**, *43*, 4534−4541.
(41) Weston, J.; Perez-Cruz, F.; Bousquet, O.; Chapelle, O.; Elisseeff, A.; Scholkopf, B. Feature selection and transduction for prediction of molecular bioactivity for drug design. *Bioinformatics* **2003**, *19*, 764−771.
(42) Müller, K.-R.; Rätsch, G.; Sonnenburg, S.; Mika, S.; Grimm, M.; Heinrich, N. Classifying "drug-likeness" with kernel-based learning methods. *J. Chem. Inf. Model.* **2005**, *45*, 249−253.
(43) Galvez, J.; Garcia-Domenech, R.; de Gregorio Alapont, C.; de Julian-Ortiz, J.; Popa, L. Pharmacological distribution diagrams: a tool for de novo drug design. *J. Mol. Graphics* **1996**, *14*, 272−276.
(44) Golland, P.; Liang, F.; Mukherjee, S.; Panchenko, D. Permutation tests for classification. *Lect. Notes Artif. Intell.* **2005**, *3559*, 501−515.
(45) Niwa, T. Using general regression and probabilistic neural networks to predict human intestinal absorption with topological descriptors derived from two-dimensional chemical structures. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 113−119.
(46) 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.* **2002**, *45*, 1737−1740.
(47) Mazzatorta, P.; Benfenati, E.; Lorenzini, P.; Vighi, M. QSAR in ecotoxicity—an overview of modern classification techniques. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 105−112.

CI050375+