Determining the Validity of a QSAR Model − A Classification Approach

Rajarshi Guha and Peter C. Jurs*

Chemistry Department, 104 Chemistry Building, Penn State University, University Park, Pennsylvania 16802

Received August 11, 2004

The determination of the validity of a QSAR model when applied to new compounds is an important concern in the field of QSAR and QSPR modeling. Various scoring techniques can be applied to specific types of models. We present a technique with which we can state whether a new compound will be well predicted by a previously built QSAR model. In this study we focus on linear regression models only, though the technique is general and could also be applied to other types of quantitative models. Our technique is based on a classification method that divides regression residuals from a previously generated model into a good class and bad class and then builds a classifier based on this division. The trained classifier is then used to determine the class of the residual for a new compound. We investigated the performance of a variety of classifiers, both linear and nonlinear. The technique was tested on two data sets from the literature and a hand built data set. The data sets selected covered both physical and biological properties and also presented the methodology with quantitative regression models of varying quality. The results indicate that this technique can determine whether a new compound will be well or poorly predicted with weighted success rates ranging from 73% to 94% for the best classifier.

INTRODUCTION

Quantitative structure−activity relationship (QSAR) modeling is based on the construction of predictive models using a set of known molecules and associated activity values. Such models can be generated using a wide variety of methods ranging from linear methods (e.g., linear regression and linear discriminant analysis) to nonlinear methods (e.g., random forests and neural networks). In all cases the predictive ability of the models are then tested with a set of molecules (the prediction set), which were not used during the model building process. Once a model has been built and validated it may be used on data for which no activity values have been measured. However, even though a model may have proved to exhibit good predictive ability based on the statistics for the prediction set, this is not always a guarantee that the model will perform well on a new set of data. The problem boils down to the fact that when a model is built and validated we have previously measured activity values to compare the predicted values with. However, when the model is applied to new data, the predicted values of the activity cannot be compared with actual values. This leads to a problem: the training set and prediction set statistics may indicate that the model has good predictive ability. But when we use the model to predict values for molecules with unknown activity, how can we be sure that the predicted activity will be close to the actual activity? If the model were able to provide some measure of confidence for its prediction, this would be helpful. Such confidence measures (also known as scores) can be defined for various models. Examples include confidence bands for linear regression models and frequency based confidence measures for decision trees. However, such measures are specific to the modeling algorithm.

This work describes a more general approach to the problem that should be applicable to any form of quantitative model. One possible approach is based on similarity. This is based on the assumption that a molecule that is structurally very similar (based on some sort of similarity metric such as atom pair similarity or fingerprint similarity) to the training set molecules will be predicted well because the model has captured features that are common to the training set molecules and is able to find them in the new molecule. On the other hand, a new molecule which has very little in common with the training set data should not be predicted very well; that is, the confidence in its prediction should be low.

An alternative approach to linking similarity measures and model quality (that is, residuals) is classification. In this method, the regression residuals for the training set are classified as good or bad, and a classification model is trained with the training set residuals. Once a trained model is obtained, we then predict the class of the prediction set residuals. However an important requirement for this process is that we be able to provide some measure of correctness for the predicted class assignments. Clearly this method does not fully solve the problem, as the classification algorithm would rarely be 100% correct. However the attractive feature of the approach discussed here is its generality. That is, it may be applied to any type of quantitative model, whether linear regression or a computational neural network. Furthermore, depending on how one defines a good residual or a bad residual, the classification model may be trained to detect unusual cases.

The fundamental decision that must be made when using the approach described in this paper is the actual class assignments of the training set residuals. Since the fact that a compound is well predicted or poorly predicted is relatively subjective (except in extreme cases), the initial assignment

* Corresponding author E-mail: pcj@psu.edu.

10.1021/ci0497511 CCC: $30.25
Published on Web 11/24/2004

of classes to the training set residuals is necessarily somewhat arbitrary. Furthermore, the nature of this class assignment defines the sizes of the two classes and hence plays a role in the choice of classification algorithm. These aspects are described in more detail in the following sections.

DATA SETS

Since one of the aims of this technique was generality, we attempted to test it on a variety of data. We considered three data sets covering both physical and biological properties. The first data set was obtained from Goll et al.[1] and consisted of the boiling points of 277 molecules obtained from the Design Institute for Physical Property Data (DIPPR) Project 801 database. This data set is relatively diverse but contains several homologous series. The second data set consisted of a 179-compound subset of artemisinin analogues described by Avery et al.[2] We have previously developed and reported linear and CNN models based on this data set.[3] The linear model from our previous study was used for the purposes of this work.

The third data set consists of 65 molecules. This data set contains 56 molecules from a study carried out by Liu et al.[4] The remaining 9 molecules were selected from the literature such that some were similar in structure to the molecules from Liu and some were distinctly different so as to be well-defined outliers in the final linear model. The molecules taken from Liu were all straight chain or branched hydrocarbons, whereas the remaining molecules included polycyclic systems as well as molecules containing heteroatoms. The dependent variable in the original work was a transformation of the boiling point defined as

$$y = \log (266.7 - \mathrm{BP})$$

where BP was the observed normal boiling points of the molecules. Since the linear model we developed for this data set did not use the all the molecules described by Liu, we did not use the logarithmic transformation and instead used boiling point values directly. The molecules and associated boiling points are shown in Table 1.

Each data set was divided into a training and prediction set. The training set was used to build a linear model, and the prediction set was used to test the models themselves as well as the algorithms developed for this study. In the case of the artemisinin data set, the same training and prediction sets that were used to develop the reported model were used in this study. Training and prediction sets for the DIPPR data set were created using the activity binning method. In both cases, the training set contained approximately 80% of the whole data set, and the remainder was placed in the prediction set. The sets for the Liu data set were created by hand. The training set contained 55 compounds selected from Liu, and the prediction set contained 10 compounds. Of these 10 compounds, one was taken from Liu, and the remaining nine were selected from the literature. The reasoning for this specific construction was to allow the prediction set to contain molecules which were very dissimilar to the training set, so that the resultant linear model would exhibit distinct outliers.

DEVELOPMENT OF LINEAR MODELS

The first step of this study involved the development of a multiple linear regression model for each data set. In the case of the artemisinin data set, we used the linear model published by Guha.[3] This model contained four descriptors, and the statistics of the model are summarized in Table 2. Linear models for the DIPPR and Liu data sets were developed using the ADAPT[5,6] methodology. A summary of the stages of this methodology is described below.

The first step involved drawing the molecular structures using Hyperchem. The resultant three-dimensional structures were geometry-optimized using MOPAC with a PM3 Hamiltonian. The next step involved the calculation of molecular descriptors. ADAPT is capable of calculating more than 250 descriptors, which fall into three main classes. The first class of descriptors are termed geometric descriptors. These characterize the three-dimensional structure of the molecule, and examples include geometric moments[7] and molecular surface areas and volumes.[8] The next class of descriptors are termed topological descriptors and consider

Table 1. Molecules and Experimental Boiling Point Values Comprising the Liu Data Set Selected by Hand from Liu et al.[4] and the Literature

name	BP (K)	name	BP (K)
methane	−164	2,2,3-trimethylpentane	110
ethane	−88.6	2,2,4-trimethylpentane	99.2
propane	−42.1	2,3,3-trimethylpentane	114.7
butane	−0.5	2,3,4-trimethylpentane	113.4
2-methylpropane	−11.7	2-methyl-3-ethylpentane	115.6
pentane	36.1	3-methyl-3-ethylpentane	118.2
2-methylbutane	27.8	2,2,3,3-tetramethylbutane	106.5
2,2-dimethylpropane	9.5	nonane	150.77
hexane	69	2-methyloctane	142.8
2-methylpentane	60.3	3-methyloctane	143.8
3-methylpentane	63.3	4-methyloctane	142.4
2,2-dimethylbutane	49.7	2,2-dimethylheptane	132.7
2,3-dimethylbutane	58	2,3-dimethylheptane	140.5
heptane	98.4	2,4-dimethylheptane	133.5
2-methylhexane	90	2,5-dimethylheptane	136
3-methylhexane	92	2,6-dimethylheptane	135.2
2,2-dimethylpentane	79.2	3,3-dimethylheptane	137.3
2,3-dimethylpentane	89.8	3,4-dimethylheptane	140.1
2,4-dimethylpentane	80.5	3,5-dimethylheptane	136
3,3-dimethylpentane	86.1	4,4-dimethylheptane	135.2
3-ethylpentane	93.5	3-ethylheptane	143
2,2,3-trimethylbutane	80.9	4-ethylheptane	141.2
octane	125.7	benzene[a]	80.1
2-methylheptane	117.6	benzoic acid[a]	249
3-methylheptane	118	cyclohexane[a]	80.7
4-methylheptane	117.7	decane[a]	174.1
2,2-dimethylhexane	106.8	bromomethane[a]	3.5
2,3-dimethylhexane	115.6	propylamine[a]	48
2,4-dimethylhexane	109.4	2,3,3-trimethylhexane[a]	131.7
2,5-dimethylhexane	109	pyrrole[a]	130
3,3-dimethylhexane	112	anthracene[a]	340
acetic acid[a]	117.9		

[a] Boiling point obtained from www.chemfinder.com.

Table 2. Statistics for the Linear Regression Model Using the Artemisinin Data Set[a]

description	beta	std. error	*t*	*P*	VIF
constant	−60.56	5.28	−11.5	2×10^{-16}	
N7CH	−0.21	0.01	−16.1	2×10^{-16}	1.6
NSB-12	0.22	0.02	9.4	2×10^{-16}	1.3
WTPT-12	27.93	2.61	10.7	2×10^{-16}	1.4
MDE−14	0.11	0.02	4.5	1.18×10^{-5}	1.5

[a] N7CH − number of 7th order chains;[9−11] NSB-12 − number of single bonds; WTPT-2 − the molecular ID number considering only carbons;[29] MDE-14 − the molecular distance edge vector considering only primary and quarternary atoms.[4]

Table 3. Statistics for the Linear Regression Model Using the DIPPR Data Set[a]

description	beta	std. error	*t*	*P*	VIF
constant	179.15628	2.03	88.32	2×10^{-16}	
FPSA-3	−175.87824	2.88	−60.95	2×10^{-16}	1.6
FNSA-3	1.36298	0.01	97.67	2×10^{-16}	1.8
RNCG-1	−0.65982	0.11	−5.65	2×10^{-16}	1.2
RPCS-1	−0.38502	0.07	−5.27	3×10^{-7}	1.1

[a] FPSA-3 − partial positive surface area divided by the total molecular surface area;[14] FNSA-3 − charge weighted partial surface area divided by the total molecular surface area;[14] RNCG-1 − the difference between the relative negative charge and the most negative charge divided by the total negative charge;[14] RPCS-1 − the positive charge analogue of RNCG-1 multiplied by the difference between relative positively charged surface area and the most positively charged surface area.[14]

Table 4. Statistics for the Linear Regression Model Using the Liu Data[a]

description	beta	std. error	*t*	*P*	VIF
constant	−381.69	60.36	−6.32	8.72×10^{-8}	
EMIN-1	−43.21	9.10	−4.75	1.95×10^{-5}	1.1
EMAX-1	88.88	10.44	8.51	4.46×10^{-11}	1.2
ECCN-1	1.27	0.11	12.08	4.49×10^{-16}	1.2
SHDW-6	501.19	136.73	3.66	6.273×10^{-4}	1.2

[a] EMIN-1 − minimum atomic estate value;[13] EMAX-2 − maximum atomic estate value;[13] ECCN-1 − eccentric connectivity index;[30] SHDW-6 − the area of the molecule when projected onto the XY plane.[31,32]

Table 5. Summary Statistics for the Three Linear Models Used in This Study

	training set		prediction set			
data set	R^2	RMSE	R^2	RMSE	*F* statistic	*P* value
artemisinin	0.70	0.87	0.05	0.75	95.28 (4156)	2.2×10^{-16}
DIPPR	0.99	7.22	0.99	7.42	9521.0 (4230)	2.2×10^{-16}
Liu	0.90	18.84	0.01	353.30	111.9 (447)	2.2×10^{-16}

the molecule as a mathematical graph. These descriptors encode various features of the resultant graphs such as connectivity information, path lengths, and various graph invariants. Essentially, they extract information regarding the shape and extent of a molecule independently of the specific geometry. Examples include χ indices,[9−11] path lengths, and electrotopological indices.[12,13] The third class of descriptors are termed electronic descriptors and represent the electronic features of a molecule such as HOMO and LUMO energies, electronegativity, and hardness. Finally, the above classes may be combined to generate hybrid descriptors. Examples include a combination of partial charges and surface areas to generate the charged partial surface area (CPSA) descriptors[14] and a combination of atomic hydrophobicity information with surface area information giving the hydrophobic surface area (HPSA) descriptors.[15,16]

Once all descriptors are calculated the pool must be reduced since many descriptors will be correlated with each other or else will be information poor. The process of descriptor reduction is carried out using two methods. First, an identical test is carried out in which descriptors for which a user specified percentage of elements (usually 85%) are zero are excluded. The second test is a correlation test which searches for pairs of descriptors having a correlation coefficient greater than a user specified value (usually 90%). From each pair, one descriptor is randomly excluded from the final pool. The size of the resultant reduced pool is generally about one-sixth the size of the actual data set.

Once a reduced pool of descriptors was generated, subjective feature selection was carried out to determine the optimal subset of descriptors to be used in a linear model. This procedure was carried out by linking a multiple linear regression routine to a simulated annealing routine[17,18] which searched the descriptor space for a descriptor subset that produced an optimal linear model. The optimality of a linear model was determined by considering the *t* statistic (models with all absolute values greater than 4.0 were accepted) as well as the overall *F* statistic. The result of this optimization procedure was a list of descriptor subsets. The top three subsets were used to build linear models, and the best model was selected by considering the overall statistics of each model. The final linear model was then subjected to a PLS analysis to ensure that the model was not overfit. The statistics of the models for the DIPPR and Liu data sets selected by this procedure are summarized in Tables 3 and 4. Summary statistics for all three linear models are presented in Table 5.

CLASSIFICATION METHODOLOGY

The aim of this work is to be able to decide whether a compound with unknown activity will be predicted well by a previously developed model. Currently, we focus only on linear regression models though the idea is general enough to be extended to other types of quantitative models (e.g., neural networks, support vector machines).

Initial attempts to develop a methodology to answer the above question focused on evaluating a similarity measure between the new compound and the training set used to develop the existing model and then attempting to correlate the similarity measure with some measure of model quality. As we restricted ourselves to linear models we considered standard error of predictions and residuals. This line of attack did lead to the observation of some general trends. That is, compounds which were more similar to the training set *generally* exhibited smaller residuals and standard error of predictions. However the observations were not conclusive, and the plots of the trends appeared to be too noisy to be able to draw any firm conclusions.

We then considered a classification approach. That is, can we classify a compound with no measured activity as well predicted or poorly predicted given a previously generated model and its associated training set? Our approach was to build a classification model using the original training set and the descriptors used in the original model and use this to predict the class of new compounds. The key word here is class. Before any model can be built we must decide how we can classify the training set. We decided to consider regression residuals, as it would allow the technique to be generalized to other types of quantitative algorithms. The training set members were classified as *bad* or *good* depending on whether their residuals were above or below a user specified cutoff value. This cutoff value plays a central role as it determines the size of the two classes. Our current strategy is to use a cutoff value obtained by visual inspection of a residual plot for the training set. The value of the cutoff

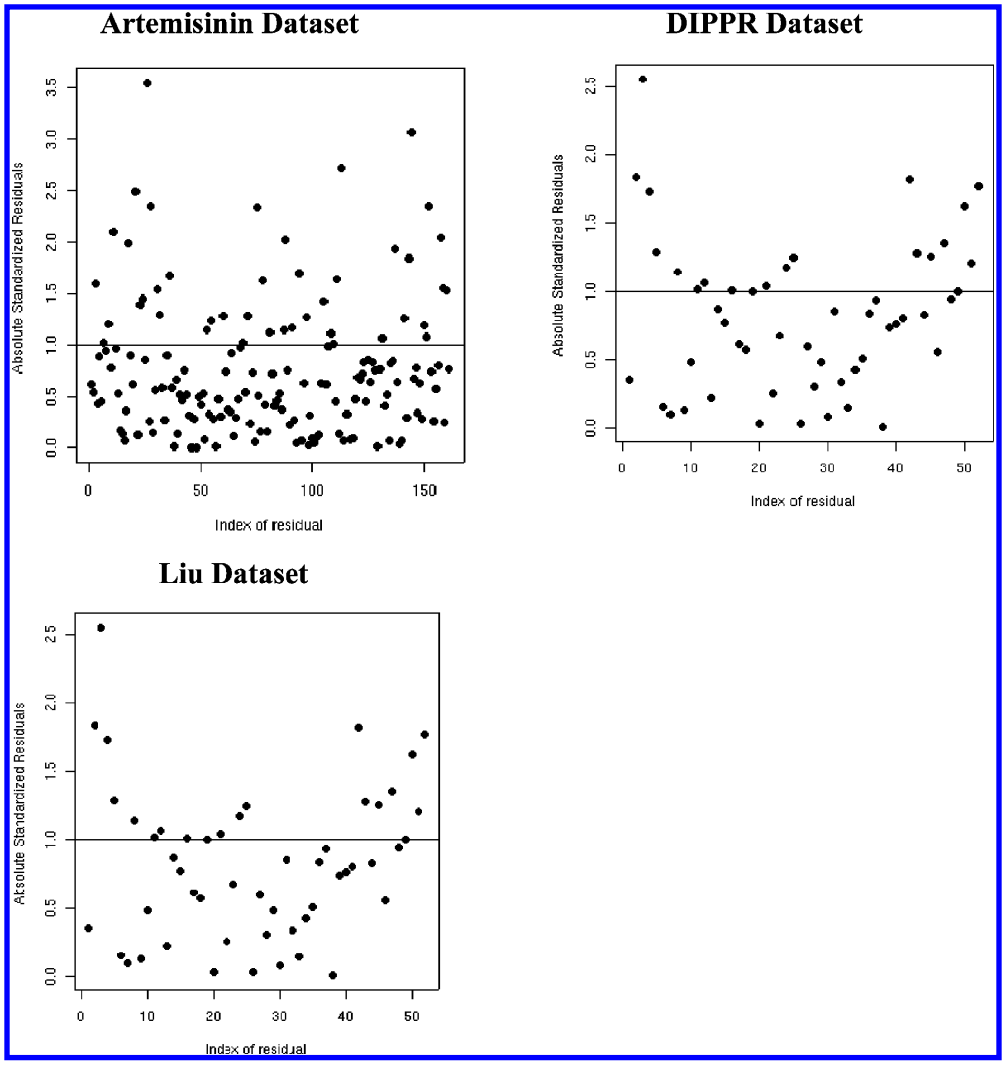

Figure 1. Plots of absolute standardized residuals versus index of residual for the best linear models developed using the training sets for each data set, with the cutoff value displayed. Residuals lying above the cutoff line are classed as *bad* and those below as *good*.

Table 6. Cutoff Values Used for Each Data Set and the Resultant Sizes of Each Class

		class size	
data set	cutoff value	good	bad
artemisinin	1.0	133	46
DIPPR	1.0	213	64
Liu	1.0	44	21

was selected so that the minor class contained approximately 20% to 30% of the whole data set. Clearly this leads to a highly imbalanced classification problem, but we felt that it would model a real world application of this technique more closely than allowing the classes to be of similar size. Alternative (nonarbitrary) methods include classifying training set members as good or bad depending on whether some regression diagnostic (Cooks distance, Mahalanobis distance) determined it was an outlier. Figure 1 shows the plots of residuals along with a line at the cutoff value for the three data sets studied here. Table 6 summarizes the cutoff values and associated class sizes for each data set.

CLASSIFICATION ALGORITHMS

Given the training set class structure, the choice of algorithm is guided by two requirements. First the goal is to be able to test compounds for which we have no measured activity value. As a result the classification algorithm must be able to produce some measure of confidence in its class predictions or else a probability of class membership (posterior probability). In the absence of such a quantity the final output of the classification model does not provide any more information than produced by simply processing the new compound through the original predictive model. The second requirement is that the algorithm must be able to handle unbalanced classes. In general, several schemes are available that can be used to modify the standard classification algorithms. These include oversampling the minority class[19] and undersampling the majority class.[20] Maloof[21] discusses the application of receiver operator characteristics (ROC) analysis in comparing how various sampling strategies and cost schemes affect classification of skewed data sets. Breiman[22] describes a simple method to increase the size of the data set without simply repeating observations. The extra samples are termed *convex pseudo data*, and the generating algorithm requires a single parameter (as opposed to kernel density methods). We investigated the use of this method in an attempt to improve classification accuracy.

We considered a wide variety of classification algorithms: logistic regression, partial least squares (PLS), discriminant analysis, neural networks, and random forests.

Table 7. Confusion Matrices for the Linear Discriminant Analysis of the Artemisinin Data Set with and without Atom Pair Similarity

training set			prediction set		
	predicted			predicted	
actual	bad	good	actual	bad	good
AP Similarity Included					
bad	2	40	bad	0	4
good	2	117	good	0	14
AP Similarity Excluded					
bad	0	42	bad	0	4
good	0	119	good	0	14

Table 8. Confusion Matrices for the Linear Discriminant Analysis of the DIPPR Data Set with and without Atom Pair Similarity

training set			prediction set		
	predicted			predicted	
actual	bad	good	actual	bad	good
AP Similarity Included					
bad	4	50	bad	1	9
good	4	177	good	1	31
AP Similarity Excluded					
bad	4	50	bad	1	9
good	4	177	good	1	31

Table 9. Confusion Matrices for the Linear Discriminant Analysis of the Liu Data Set with and without Atom Pair Similarity

training set			prediction set		
	predicted			predicted	
actual	bad	good	actual	bad	good
AP Similarity Included					
bad	7	11	bad	2	1
good	4	30	good	3	7
AP Similarity Excluded					
bad	4	14	bad	3	0
good	4	30	good	2	8

Random forests were first described by Breiman[23] and have been used in a variety of QSAR applications. The original random forest algorithm was not suited for very unbalanced data sets, but current implementations[24] use a weighting scheme which overcomes this problem. We decided not to use this algorithm, due to the fact that it works with large descriptor pools. This capability is due to its ability to ignore irrelevant descriptors as well as the fact the algorithm is resistant to overfitting. We did not want to build the classification models with more (and different) information than was available to the original regression models. Since part of the performance for random forest models is due to its ability to build trees based on good descriptor subsets, restricting the descriptor pool to four or five descriptors would probably result in lower quality random forest models.

In the case of discriminant analysis, we investigated the use of linear and quadratic discriminant methods. In each case, the algorithms employed were able to generate posterior probabilities via a cross-validation scheme. As the results were quite similar we only present the results of the linear discriminant analysis. All the algorithms mentioned above were obtained from the R software package.[25]

As mentioned in the previous section we used the descriptors from the original regression model to build the classification model. We also investigated the use of a similarity measure as a source of extra information. Intuitively, one would expect that a new molecule that is similar to the molecules in the training set should be well predicted by the regression model. Thus, in addition to classification models built only with descriptors from the regression model we also built models that also contained a similarity value. We chose to use the atom pair similarity described by Carhart.[26] Atom pair similarities are calculated between pairs of molecules. To provide a single similarity value for each compound we calculated the average similarity value between each compound and all the compounds in the training set.

RESULTS

Most of the algorithms exhibited good predictive ability considering the fact that the data sets used were not very large (especially the Liu data set). As expected, the neural network performed very well, with a 90% correct prediction rate on the training set and 72% to 85% correct on the prediction set. The inclusion of the similarity values as a descriptor did not appear to improve the results significantly.

Table 7 shows the confusion matrices for the training and prediction sets generated using linear discriminant analysis for the artemisinin data set. The implementation used for this study allowed us to specify the prior probabilities for each class. We assumed that the priors could be approximated by the class proportions. Clearly, the very poor predictions for the minority class indicate the problem due to the imbalanced nature of the class distributions. To try and remove the bias due to the imbalanced nature of the problem, the model was regenerated with an oversampled minority class. However the results did not improve significantly. To investigate whether extra information might improve the situation we also regenerated the model using the averaged atom pair similarity values as an extra independent variable. We felt that this was justified (as compared to using extra molecular descriptors) since this descriptor essentially compares the molecules among themselves. Table 7 displays the confusion matrices for the model. The predictions for the good class are now 100%, but members of the bad class are mispredicted in all cases. The results for this algorithm when applied to the DIPPR data set give similar results. The classes assigned in this data set are also quite unbalanced. The confusion matrices are presented in Table 8. The results for the Liu data set (Table 9) are marginally better, more so for the prediction set than the training set. This is probably due to the slightly higher proportion of the minor class in the training set.

The results from the PLS classification scheme were not significantly better than those obtained with LDA and in some cases worse, and as a result we omit their presentation.

Tables 10−12 present the confusion matrices for the three data sets generated using a neural network. The network used entropy outputs[27] and thus provided the associated probabilities with each class assignment. In all cases, the inverse of the class proportions were used as example weights. Table 10 shows that the performance for the artemisinin data set was not very impressive. However the imbalanced nature of the data set does not affect the performance as much as in the case of LDA. In contrast, the DIPPR data set showed a very good performance using the neural network methodology as can be seen from Table 11. In this case, the bad class was very well predicted in both the training and

Table 10. Confusion Matrices Generated by the Neural Network Applied to Artemisinin Data Set with and without Atom Pair Similarity[a]

training set			prediction set		
	predicted			predicted	
actual	bad	good	actual	bad	good
AP Similarity Included					
bad	38	4	bad	4	0
good	27	92	good	3	11
AP Similarity Excluded					
bad	34	8	bad	3	1
good	46	73	good	4	10

[a] The architecture of the CNN with atom pair similarity excluded was 4−9−1 and with similarity included was 5−5−1.

Table 11. Confusion Matrices Generated by the Neural Network Applied to the DIPPR Data Set with and without Atom Pair Similarity[a]

training set			prediction set		
	predicted			predicted	
actual	bad	good	actual	bad	good
AP Similarity Included					
bad	54	0	bad	9	1
good	5	176	good	1	31
AP Similarity Excluded					
bad	54	0	bad	8	2
good	5	176	good	2	30

[a] The architecture of the CNN with atom pair similarity excluded was 4−5−1 and with similarity included was 5−4−1.

Table 12. Confusion Matrices Generated by the Neural Network Applied to the Liu Data Set with and without Atom Pair Similarity[a]

training set			prediction set		
	predicted			predicted	
actual	bad	good	actual	bad	good
AP Similarity Included					
bad	18	0	bad	3	0
good	1	33	good	2	8
AP Similarity Excluded					
bad	17	1	bad	30	0
good	2	32	good	2	8

[a] The architecture of the CNN with atom pair similarity excluded was 4−5−1 and with similarity included was 5−5−1.

prediction sets. Finally the Liu data set also yielded good results (Table 12). In all cases, the use of average atom pair similarity as an extra independent variable did not appear to improve results.

Table 13 displays the weighted success rates for all the classification methods on all the data sets. This measure of classification success was described by Weston et al.[28] and is defined as

$$w = \frac{1}{2}\left(\frac{\text{no. true positive}}{\text{total positives}}\right) + \frac{1}{2}\left(\frac{\text{N true negatives}}{\text{total negatives}}\right)$$

The above expression indicates that $0 \leq w \leq 1$. As mentioned by Weston, this measure is suitable for unbalanced classification problems. The values indicate the poor performance of the LDA (in fact it appears to be not much better than random) and PLS methods and the much better performance of the neural network approach.

Table 13. Weighted Success Rates for the Various Classification Algorithms

		without similarity		with similarity	
method	data set	TSET	PSET	TSET	PSET
LDA	artemisnin	0.51	0.50	0.50	0.50
	DIPPR	0.52	0.53	0.52	0.53
	Liu	0.63	0.68	0.55	0.90
PLS	artemisnin	0.51	0.46	0.49	0.50
	DIPPR	0.36	0.53	0.36	0.53
	Liu	0.59	0.51	0.59	0.73
CNN	artemisnin	0.79	0.80	0.71	0.73
	DIPPR	0.98	0.93	0.98	0.86
	Liu	0.98	0.90	0.94	0.90

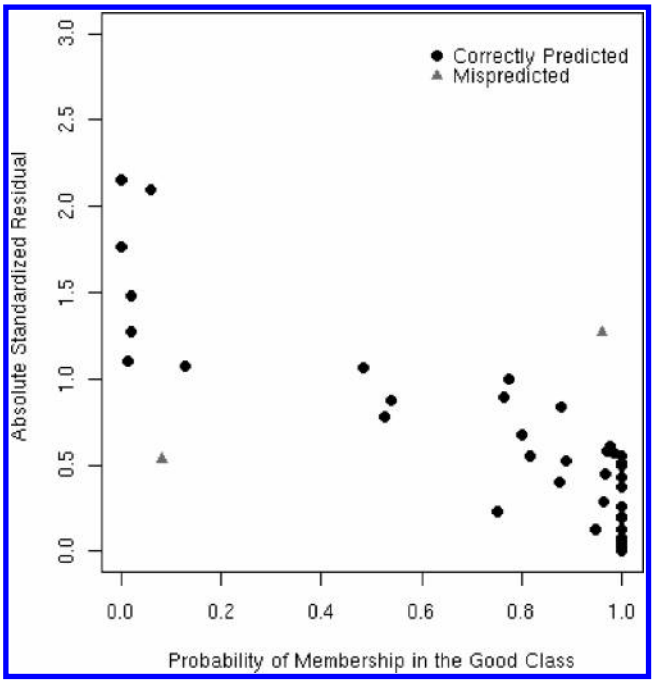

Figure 2. Plot of probability of membership to the good class versus the absolute standardized residual for the DIPPR data set. The probabilities were obtained from the CNN model. The prediction set portion was used to generate the plot.

We also attempted to improve the classification results by using the convex pseudo data method described by Breiman[22] to increase the training set sizes. We considered two approaches. In the first method, we simply extended the whole data set without regard to class. The new samples were placed in the training set, and the extended training sets were used to build models. In the second approach we only extended the portion of the training set that was assigned to the bad class (essentially increasing the size of the bad class). However, though the results in some cases (PLS and LDA) did improve to some extent, the increases in classification rates did not appear to be significant, and hence we omit them in this study.

One way to consider the performance of the models is shown in Figures 2−4. The probability for membership in the good class is plotted against the residuals from the original linear regression model. The probabilities were obtained from the neural network classification models. Figure 2 is the plot for the prediction set of the DIPPR data set. Ideally one would expect that such a graph would have a cluster of points in the upper left quadrant and a cluster in the lower right quadrant. However, in practice such a perfect distribution is rare, although the graph does indicate the

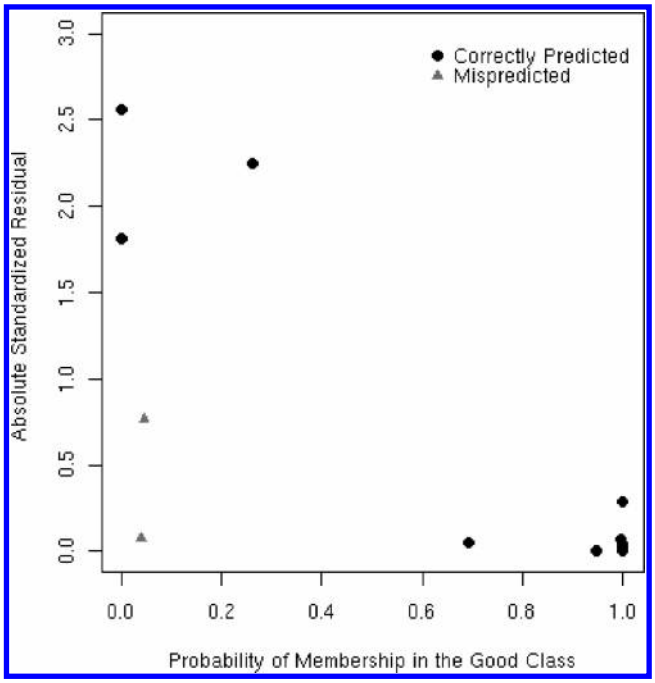

Figure 3. Plot of probability of membership to the good class versus the absolute standardized residual for the Liu data set. The probabilities were obtained from the CNN model. The prediction set portion was used to generate the plot.

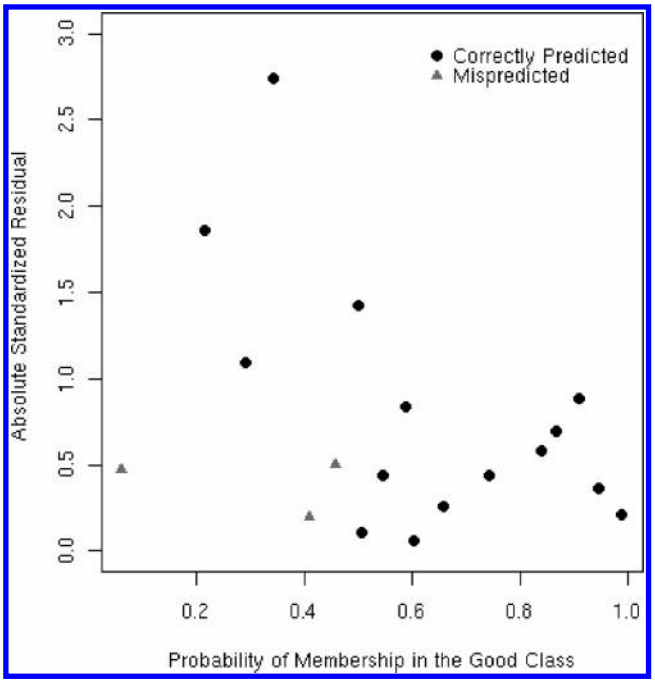

Figure 4. Plot of probability of membership to the good class versus the absolute standardized residual for the artemisinin data set. The probabilities were obtained from the CNN model. The prediction set portion was used to generate the plot.

general trends. In the lower right there is a vertical set of points with probability 1.0 that exhibit low values of the absolute standardized residual. On the left-hand side of the graph, we see a similar set of points with probability values equal to 0.0 (indicating that they belong to the bad class). Between these two extremes we see points that have probabilities indicating membership to the good class. However for points with probabilities lying in the range 0.5 to 0.7 such membership is probably not conclusive, and we see that their residuals are also midway between the two extremes. The points at the left and right edges of the graph indicate that when the class predictions of the CNN classifier are accompanied by high or low probabilities, the residuals from the linear regression model can be expected to be low or high, respectively. The two anomalous points marked by red triangles represent the misclassified cases. The one on the right was predicted as belonging to the good class, whereas its true membership was to the bad class and vice versa for the point on the right. It is not apparent why these points would be misclassified. But more importantly, it is not clear how one might consider them misclassified without having the actual residuals available, since in a real application we would be dealing with observations whose actual activities are not known.

Figure 3 shows the corresponding plot for the Liu data set. As before, observations predicted to be in the bad class and the good class (with high certainty) are located in the upper left and lower right quadrants, respectively. In this case, there is only a single point whose membership is not absolutely certain.

Finally, Figure 4 shows the plot for the artemisinin data set. In this case the plot is not as tight as the previous ones with the probability values of a number of observations indicating that membership in the good class is not very conclusive. The misclassified observations are interesting. The one misclassified point on the left-hand edge would certainly be difficult to detect in the absence of residuals. However, the remaining two misclassified points are more or less on the border between the two lower quadrants. In addition they are also quite close to points that have been correctly classified. This is indicative of the fact that membership of observations when their probabilities lie around 0.5 can be inconclusive, and thus one should be wary of such points.

FURTHER WORK

The methodology described above appears to perform reasonably well on the three data sets we investigated. However, there are several features that require further study. First, the classification approach described here is a two-class problem. We restricted ourselves to the two-class problem for simplicity. However, considering the scheme as a three-class problem might enable the user to draw more fine-grained conclusions regarding the validity of the results obtained from a regression model. However, increasing the number of classes will certainly require a large data set and even if such a data set is used, the unbalanced nature of the classes will require careful selection of a classification technique. We note that the results presented in this study are dependent on the nature of the data sets employed—specifically the distribution of residuals which is itself dependent on the distribution of the compounds in descriptor space. However, the data sets that we selected for testing include both physical properties for a number of congeneric series as well as biological properties for a set of molecules containing exhibiting varying structures and functionality. Furthermore the data sets we selected allowed us to test our techniques with different types of linear models. For example, the DIPPR data set was described by a linear model with very good statistics and very low residual values in general. On the other hand, the artemisinin data set was characterized

by lower values of R^2, high RMSE value, and a number of observations with large residuals. As a result the DIPPR data set presented our methodology with severely unbalanced classes, whereas the class distribution was not as skewed in the case of the arteminsin data set. Furthermore it is often the case that linear models for biological properties do not exhibit high quality statistics and contain a number of outliers. Thus the use of this data set allowed us to test our technique in a real world scenario. Finally, the Liu data set that was prepared by hand allowed us to have specific observations with large residuals and thus test the ability of the methodology to specifically detect these types of compounds. As has been shown, our methodology appears to perform well on these varied data sets. The only downside to the selection of our data sets is that the sizes are not as large as we would have liked them to be. Larger data sets would allow us to experiment with more than two classes as well as other classification schemes as discussed below. Clearly, one possible avenue of investigation is the validation of our methodology on different (and larger) data sets.

Modified sampling schemes such as those described do not appear to improve the results significantly. The initial assignment of classes to the training set data is a step that could be modified as the current work describes an arbitrary assignment scheme. To remove this user defined task, class assignments can be automated by the use of regression diagnostics. However, such a scheme would then restrict the application of this methodology to linear models only. It appears that for full generality some form of cutoff value must be specified by the user. However, one advantage of a user-specified cutoff value is that it allows the user to focus on a range of residual values. Coupled with multiple (more than two) classes, this would allow the user to perform a fine-grained analysis of the residual classes.

Of the classification techniques investigated in this study it appears that neural networks performed the best with overall classification rates ranging from 79% to above 90% for the training set and 73% to 90% for the prediction set. The linear methods did not appear to perform significantly better than random. Furthermore, introduction of a similarity measure as an independent variable did not lead to improved classification results using any of the methods.

An alternative approach that may be considered is a Bayesian classification scheme whereby the training set class assignments are used to build up a prior probability distribution and the probability of new compounds belonging to a given class can be obtained by sampling from the simulated distribution. Associated with each class prediction is a probability for the membership to the predicted class. This requirement restricted our choice of classification techniques somewhat, but we feel that the lack of such a posterior probability would result in this method not being any more useful than simply recalculating the original regression model with some sort of scoring feature. The plots of posterior probability versus residuals are a good indicator of the performance of this methodology and also allows us to identify misclassifications in general. However, misclassified examples that are associated with posterior probabilities around 0.5 are, in general, not distinguishable from correctly predicted examples with similar posterior probabilities. In such cases one would probably be justified in ignoring compounds whose class predictions are borderline and rather concentrate on those compounds that are classified with high posterior probabilities of belonging to the good or bad class.

CONCLUSIONS

The work presented here describes a novel and general scheme to provide a measure of confidence for the predictions from a regression model. The methodology described here attempts to answer the following question: how well will a regression model predict the property value for a compound that was not in the training or prediction set of the model. Multiple approaches were investigated resulting in a classification scheme in which the training set residuals were assigned to one of two classes depending on whether they lay above or below a cutoff value. A classifier was then built with these assignments and used to predict the class of the residual for a new compound. The technique appears to be general enough to be applicable to any given regression model. We investigated several classification techniques, and a neural network approach produced the best classification rates. The performance of the algorithm was visualized by considering plots of posterior probabilities versus residuals.

Though the performance of regression models may be judged via other scoring methods, such as confidence bands or frequency based scores, these methods are generally specific to the regression modeling technique employed. Our method is quite general and thus can be applied to regression models developed using linear regression, neural networks, or random forests. Furthermore, our methodology is not dependent on the original data set. All that is required is the availability of the original residuals (which is generally available in models developed with common statistical packages). Another attractive feature is that apart from the threshold residual value, the methodology does not require extra information such as similarity measures or new descriptors, since it restricts itself to using the descriptors that were used in the original quantitative model. We believe that such a parsimonious approach minimizes complexity as well as user intervention. The net result of our methodology is a probability of whether a compound (with an unknown property value) will have a high or low residual (relative to a user specified cutoff value) when processed by the regression model. Clearly, this does not replace the use of the original quantitative model. Rather, the methodology allows us to generate confidence measures for new compounds for any type of quantitative regression model in the absence of the original data and in a parsimonious manner. As a result methodology could be used as a component of a high throughput screening process in which different regression techniques are employed in a consensus based strategy.

ACKNOWLEDGMENT

This work was partially supported by NSF grant No. CHE-0333222. We would also like to thank Dr. Brian Mattioni for numerous helpful discussions during this study.

REFERENCES AND NOTES

(1) Goll, E. S.; Jurs, P. C. Prediction of the normal boiling points of organic compounds from molecular structures with a computational neural network model. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 974−983.

(2) Avery, M. A.; Alvim-Gaston, M.; Rodrigues, C. R.; Barreiro, E. J.; Cohen, F. E.; Sabnis, Y. A.; Woolfrey, J. R. Structure activity relationships of the antimalarial agent artemisinin. the development of predictive in vitro potency models using comfa and hqsar methodologies. *J. Med. Chem.* **2002**, *45*, 292−303.
(3) Guha, R.; Jurs, P. C. The development of QSAR models to predict and interpret the biological activity of artemisinin analogues. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1440−1449.
(4) Liu, S.; Cao, C.; Li, Z. Approach to estimation and prediction for normal boiling point (nbp) of alkanes based on a novel molecular distance edge (mde) vector, lambda. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 387−394.
(5) Jurs, P. C.; Chou, J. T.; Yuan, M. Studies of chemical structure biological activity relations using patter recognition. In *Computer assisted drug design*; Olsen, E. C., Christoffersen, R. E., Eds.; American Chemical Society: Washington, DC, 1979.
(6) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer assisted studies of chemical structure and biological function*; Wiley: New York, 1979.
(7) Goldstein, H. *Classical mechanics*; Addison-Wesley: Reading, MA, 1950.
(8) Pearlman, R. S. *Physical chemical properties of drugs*; Yalkowsky, S. L., Sinkula, A. A., Valvani, S. C., Eds.; Marcel Dekker, Inc.: New York, 1980.
(9) Kier, L. B.; Hall, L. H.; Murray, W. J. Molecular connectivity i: relationship to local anesthesia. *J. Pharm. Sci.* **1975**, *64*,
(10) Kier, L. B.; Hall, L. H. *Molecular connectivity in structure activity analysis*; John Wiley & Sons: 1986.
(11) Kier, L. B.; Hall, L. H. Molecular connectivity vii: specific treatment to heteroatoms. *J. Pharm. Sci.* **1976**, *65*, 1806−1809.
(12) Kier, L. B.; Hall, L. H. *Molecular structure description. the electrotopological state*; Academic Press: London, 1999.
(13) Kier, L. B.; Hall, L. H. An electrotopological-state index for atoms in molecules. *Pharm. Res.* **1990**, *7*, 801−807.
(14) Stanton, D. T.; Jurs, P. C. Development and use of charged partial surface area structural descriptors in computer assissted quantitative structure property relationship studies. *Anal. Chem.* **1990**, *62*, 2323−2329.
(15) Stanton, D. T.; Mattioni, B. E.; Knittel, J. J.; Jurs, P. C. development and use of hydrophobic surface area (hsa) descriptors for computer-assisted quantitative structure−activity and structure−property relationship studies. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1010−1023.
(16) Mattioni, B. E. *The development of qsar models for physical property and biological activity prediction of organic compounds*; 2003.
(17) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M.; Teller, A. H.; Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087−1092.
(18) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated descriptor selection for quantitative structure−activity relationships using generalized simulated annealing. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 77−84.
(19) Japkowicz, N. Learning from imbalanced data sets: a comparison of various strategies. In *Learning from imbalanced data sets: papers from the aaai workshop*; AAAI Press: 2000.
(20) Kubat, M.; Matwin, S. Addressing the curse of imbalanced training sets: one sided selection. In *Proceedings of the fourteenth international conference on machine learning*; Morgan Kauffmnan: 1997.
(21) Maloof, M. A. Learning when data sets are imbalanced and when costs are unequal and unknown. In *ICML-2003 Workshop on Learning from Imbalanced Data Sets II*; 2003.
(22) Breiman, L. *Using convex pseudo data to increase prediction accuracy*; 1998.
(23) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and regresion trees*; Bickel, P., Cleveland, W. S., Dudley, R. M., Eds.; Wadsworth International Group: 1984.
(24) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, C.; Sheridan, R. P.; Feuston, B. P. Random forest: a classification and regression tool for compound classification and qsar modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *42*, 1947−1958.
(25) R Development Core Team. R: a language and environment for statistical computing. 2004, http://www. R-project.org.
(26) Carhart, R. E.; Smith, D. E.; Venkataraghavan, R. Atom pairs as molecular features in structure−activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.
(27) Ripley, B. D. *Pattern Recognition and Neural Networks*; Cambridge University Press: Oxford, 1996.
(28) Weston, J.; Pérez-Cruz, F.; Bousquet, O.; Chapelle, O.; Elisseeff, A.; Schölkopf Feature selection and transduction for prediction of molecular bioactivity for drug design. *Bioinformatics* **2003**, *19*, 764−771.
(29) Randic, M. On molecular idenitification numbers. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 164−175.
(30) Sharma, V.; Goswami, A.; Madan, A. K. Eccentric connectivity index: a novel highly discriminating topological descriptor for structure−property and structure−activity studies. *J. Chem. Inf. Comput. Sci.* **1998**, *37*, 273−282.
(31) Stouch, T. R.; Jurs, P. C. A simple method for the representation, quantification and comparison of the volumes and shapes of chemical compounds. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 4−12.
(32) Rohrbaugh, R. H.; Jurs, P. C. Molecular shape and prediction of high performace liquid chromatographic retention indices of polycyclic aromatic hydrocarbons. *Anal. Chem.* **1987**, *59*, 1048.

CI0497511