Journal of Molecular Graphics and Modelling 35 (2012) 21–27

Contents lists available at SciVerse ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

In silico prediction of toxic action mechanisms of phenols for imbalanced data with Random Forest learner

Jing Chen [a,*], Yuan Yan Tang [a,b], Bin Fang [a], Chang Guo [a]

[a] College of Computer Science, Chongqing University, Chongqing 400030, China
[b] Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Taipa, Macau, China

ARTICLE INFO

Article history:
Received 6 August 2011
Received in revised form 7 January 2012
Accepted 9 January 2012
Available online 17 January 2012

Keywords:
Toxic action mechanisms
Phenols
QSAR
Random Forest
Cost-sensitive

ABSTRACT

With an increasing need for the rapid and effective safety assessment of compounds in industrial and civil-use products, *in silico* toxicity exploration techniques provide an economic way for environmental hazard assessment. The previous *in silico* researches have developed many quantitative structure–activity relationships models to predict toxicity mechanisms for last decade. Most of these methods benefit from data analysis and machine learning techniques, which rely heavily on the characteristics of data sets. For *Tetrahymena pyriformis* toxicity data sets, there is a great technical challenge—data imbalance. The skewness of data class distribution would greatly deteriorate the prediction performance on rare classes. Most of the previous researches for phenol mechanisms of toxic action prediction did not consider this practical problem. In this work, we dealt with the problem by considering the difference between the two types of misclassifications. Random Forest learner was employed in cost-sensitive learning framework to construct prediction models based on selected molecular descriptors. In computational experiments, both the global and local models obtained appreciable overall prediction accuracies. Particularly, the performance on rare classes was indeed promoted. Moreover, for practical usage of these models, the balance of the two misclassifications can be adjusted by using different cost matrices according to the application goals.

1. Introduction

Phenols have been commonly used in industrial chemicals and in consumer products, such as stabilizers, antioxidants, dyestuffs, detergents, pesticides, explosives, plastics and drugs [1]. Due to their long term stability under the ordinary condition and their toxicity characteristics, they are significantly responsible for environment pollutions and a variety of acute and chronic diseases, such as mutagenicity, carcinogenicity, and dysgenesis [2,3]. The evaluation of phenol toxicity has become an area of active research in the past decades. Recently, *in silico* methods, such as quantitative structure–activity relationships (QSAR) and quantitative structure–toxicity relationships (QSTR), bring many benefits for toxicity exploration [4,5]. On the one hand, the high cost and time needed to conduct the safety testing for a large volume of chemicals are heavily reduced by *in silico* methods. It promotes our ability to meet the commercial requirement. On the other hand, the usage of animals in ecotoxicological risk assessment can be reduced as required under some legislations and the humanism.

Several QSAR and QSTR models were developed to explore phenol toxicity in the past decade [6–8]. In particular, the toxicity evaluation of *Tetrahymena pyriformis* is an attractive research topic for the environment impact assessment of toxicants [9]. To explore toxicity of phenols according to mechanisms of toxic action, some used qualitative approaches based on simple structural characteristics (such as the presence of a certain substituent) [10]. It is restricted by the limited types of substituents in the training set. Other researchers consider intelligent data analysis methods which are commonly used in artificial intelligence area. These methods identify toxic action mechanisms of phenols in terms of physicochemical properties. There are still some disadvantages, of which a significant one is that the classification performance of these methods is heavily dependent on the characteristics of training dataset, such as class distribution, data size, and noise. In this work, intelligent data analysis methods were considered.

A variety of intelligent methods were utilized to develop QSAR models, including linear discriminant analysis [11], artificial neural networks [12], multilinear regression [9], decision tree analysis [13], AdaBoost learner [14], support vector machine [15], ensemble learning models [16] and so on. From the reported results, these models archived appreciable performance. However, it is known

* Corresponding author at: Room 1718, Main Building, Campus A, Chongqing University, Chongqing 400030, China. Tel.: +86 13628466455.
E-mail addresses: chenjingmc@gmail.com (J. Chen), yytang@umac.mo (Y.Y. Tang), fb@cqu.edu.cn (B. Fang), changguocqu@gmail.com (C. Guo).

1093-3263/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2012.01.002

Table 1
Summarization of several TPT data sets.

Data sets	Year	# Compounds	# Descriptors	# Classes	Class distribution
(a) Russom et al. [19]	1997	408	5	6	239:96:36:16:12:9
(b) Cronin et al. [18]	2002	250	108	5	173:27:27:19:4
(c) Aptula et al. [11]	2002	221	5	4	153:27:23:18
(d) Xue et al. [20]	2006	1129	199	2	841:288
(e) Cheng et al. [15]	2011	1571	166	2	1217:354

that the results of data analysis methods are heavily dependent on the characteristics of data sets. For the construction and evaluation of these models, a high-quality data set is necessary. Jalali-Heravi and Kyani reported a data set of 268 substituted benzenes to *T. pyriformis* toxicity (TPT) [12]. In this data set, five mechanistically interpretable molecular descriptors were used to represent a phenolic compound. Cronin et al. constructed a biological data set of 250 phenols, some of which was previously reported [17]. Cronin et al. divided the data set into two partitions: the one with 200 phenols to TPT for models building and the other with 50 phenols for the validation of QSAR models. For each compound, a total of 108 physico-chemical descriptors were calculated. Aptula et al. developed a linear discriminant analysis based QSAR model for 221 phenols using some simple descriptors such as $\log K_{ow}$, pK_a, E_{LUMO}, E_{HOMO} and N_{Hdon} [11]. For this data set, later, additional molecular descriptors were calculated and the enlarged version data set of 220 phenols to TPT was constructed [16].

These data sets share a challenging characteristic—data imbalance. Several data sets of phenolic compounds to TPT were summarized in Table 1. These data sets had a wide range of total sizes, descriptors, and classes. Class distributions of these data sets are all seriously skewed (Fig. 1). Particularly, the most dramatical ratio of the majority to the minority is high to 43:1 in the data set Cronin et al. used [18]. For the binary classification problem (TPT and non-TPT), there are often more TPT compounds than non-TPT compounds. For multi-mechanism classification, "polar narcotics" are usually in majority. The compounds with other mechanisms are relatively fewer. In these data sets, the majority class is over-represented, while the minority classes cannot be described sufficiently. It resulted in a direct problem that the constructed QSAR models usually have skewed performance on skewed classes. Many techniques have been developed in data mining and machine learning field to handle this problem. However, most of TPT QSAR prediction models did not consider the imbalanced data problem [11,13,16], which often occurred in TPT datasets. Thus those QSAR models may achieve appreciable accuracy on the majority class and overall data set, but perform poorly on rare classes.

The goal of this study was to promote the discriminative ability on minority classes. Furthermore, the balance of the model performances on both majority classes and minority classes was exploring. An effective machine learning tool—Random Forest—was used for mechanisms classification on imbalanced TPT data set. We made the original Random Forest cost-sensitive for imbalanced data by both sampling and thresholding. Correlation-based feature selection algorithm with Best First search was applied to optimal descriptors selection. From the *in silico* experimental results, the prediction performance was greatly improved after cost-sensitive learning framework was employed.

2. Materials and methods

2.1. Biological data

The phenol data set was obtained from the reference [11], in which toxicity data of the ciliate *T. pyriformis* for a set of 221 phenols were available. Each phenol was assigned with one of four toxic action mechanisms: polar narcotics, weak acid respiratory uncouplers, pro-electrophiles and soft electrophiles. It was done by using simple structural rules developed earlier for the growth inhibition assay with *T. pyriformis* [17]. The class distribution of this data set was summarized in Table 2. The data was heavily skewed.

In the literature [11], all phenols were divided into two equalized complementary subsets: *group* 1 and *group* 2, with respect to training and testing. In this work, the same partition strategy was adopted. Firstly, *group* 1 was used as training set to construct the QSAR prediction models, which were evaluated by using the testing set *group* 2. And then *group* 2 was used to build the models, which were evaluated with *group* 1 as testing set. Finally, the two sets of results were numerically averaged as the final results to be reported.

The original molecular descriptors used in the reference [11] included $\log K_{ow}$, pK_a, E_{LUMO}, E_{HOMO} and N_{Hdon}. Norinder et al. extended the molecular expression by adding more descriptors [16]. In the work of Norinder et al. [16], more molecular descriptors were calculated by using the SELMA program [21]. The final total 51 molecular descriptors were used to represent each compound. All phenols were labeled by toxic action mechanisms as that in the literature [11]. So the skewed class distribution was the same as the case shown in Table 2. In this study, we used this extended data set as our experimental data. The selected molecular descriptors were the same as those in the literature [16].

2.2. Machine learning methods

2.2.1. Random Forest

Random Forest [22] is a popular classification algorithm. Several bootstrap sample sets of the training data are constructed by using random feature selection procedure. Each sample set with the random selected descriptors is used to induct an unpruned classification tree. The final decision is made by majority voting to aggregate the predictions of the ensemble of decision trees. In the previous work [22], Random Forest generally shows a significant performance improvement over the single tree classifiers such as CART [23] and C4.5 [13]. Particularly, it is more robust to noise than AdaBoost [14], which is another popular ensemble learner. However, similar to other common classifiers, Random Forest suffers from the extremely imbalanced data problem. Since it is trained to minimize an overall error rate—a percentage of incorrect predictions, it will tend to focus on prediction accuracy of majority class, which often results in poor accuracy for minority class. To alleviate the problem, in this study we introduce cost-sensitive learning with Random Forest as its base classifier.

Table 2
Class distribution of the data set.

MOA class	*Group* 1	*Group* 2	Total	Percentage
Polar narcotics	75	77	152	69.1%
Weak acid respiratory uncouplers	9	9	18	8.2%
Precursors to soft electrophiles	13	14	27	12.3%
Soft electrophiles	11	12	23	10.5%
Total	108	112	220	100.0%

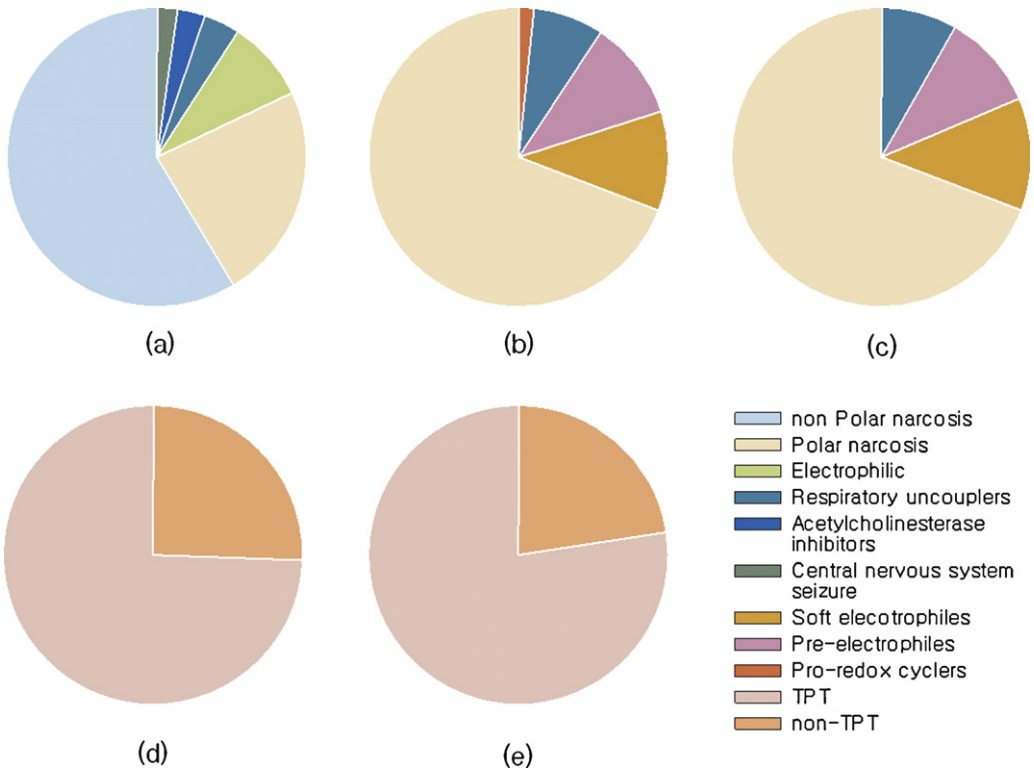

Fig. 1. Class distributions of the data sets in Table 1.

2.2.2. Cost-sensitive classification

Cost-sensitive classification is a technique to handle imbalanced data problem. Without the loss of generality, we considered a binary classification at first. The case was easy to be extended into multi-class problem. There were two types of misclassifications for a binary case. The first mistake was to misclassify a majority class compound as a minority class one. The inverse was the second mistake. Most classification methods ignored the difference between the two types of misclassification errors. Especially, they were implicitly based on the assumption that the costs of all errors were the same. In fact, the model trained by imbalanced data will tend to make the first type of mistakes. In this study, we considered the different misclassification costs, which were presented in the cost matrix (Table 3).

CostSensitiveClassifier (CSC in short) was introduced in the reference [24], and it is commonly used for imbalanced data. CSC makes its base classifier cost-sensitive by using two basic modes: sampling and thresholding.

In cost matrix, the notation $C(i, j)$ presents the misclassification cost to classify an instance from its actual class j into predicted class i. Usually, we assume that correctly classifying an instance x from class i into class i has no cost. Supposing the class labels were 0 (negative class) and 1 (positive class), we had $C(0, 0) = C(1, 1) = 0$. The mistake cost values can be obtained from domain experts, or learned by other approaches. In this study, the cost matrices were adjusted to be appropriate during training process by hand.

Table 3
The cost matrix of binary classification.

	Predicted negative	Predicted positive
Actual negative	$C(0, 0)$	$C(1, 0)$
Actual positive	$C(0, 1)$	$C(1, 1)$

Given a cost matrix, thresholding-based CSC pursues a minimum expected overall cost. Considering an instance x, the expected cost of classifying x into class i can be expressed as:

$$R(i|x) = \sum_j P(j|x)C(j, i) \quad (1)$$

where $P(j|x)$ is the probability estimation of classifying x into class j. Then an instance x can be classified into the positive class, if and only if:

$$R(0|x)C(1, 0) + P(1|x)C(1, 1) \leq P(0|x)C(0, 0) + P(1|x)C(0, 1) \quad (2)$$

As $C(0, 0) = C(1, 1) = 0$, so the criterion of classifying an instance as the positive class becomes as

$$R(0|x)C(1, 0) \leq P(1|x)C(0, 1) \quad (3)$$

Substituting $P(0 \mid x) = 1 - P(1 \mid x)$ into the above equation, we obtained p^* as the threshold to classify an instance x into the positive class if $P(1 \mid x) \geq p^*$.

$$p^* = \frac{C(1, 0)}{C(1, 0) + C(0, 1)} \quad (4)$$

Similarly, for a multi-class problem, the thresholding-based CSC computes the probability estimations $P(j \mid x)$ of each test instance, and uses (1) to predict the class label of each test instance in terms of the minimum expected cost principle.

The other mode of CSC algorithm uses sampling technique to be made cost-sensitive. The sampling-based CSC is to re-weight the data according to the cost matrix so that the generated model is inherently less likely to make expensive mistakes. Considering a specific rare class i, we assigned a much larger weight to the mistake of incorrectly classifying instances of class i. This was done by sampling the data so that the class i was over represented, since an instance from class i had a higher probability to be chosen into

the new sample set. After a new sample set was constructed, the original Random Forest was performed on it.

2.3. Performance evaluation

This section provided a brief description of measures used in performance evaluation. These measures were adopted, since they are commonly used in class imbalance problem. We only considered the explanation of these measures for binary problem, namely, for local prediction models. It was easy to be extended for the 4-mechanism global problem. For a considered class, we called it a positive class, and the other was negative.

These measures were all computed from confusion matrices. *TP* and *FN* were the numbers of correctly and incorrectly classified compounds of the actual positive class, respectively. Similarly, *TN* and *FP* denoted the numbers of correctly and incorrectly classified compounds of the actual negative class. Then we could construct the confusion matrix as [*TN*, *FP*; *FN*, *TP*].

2.3.1. Accuracy

Accuracy is defined as the ratio of the number of compounds correctly classified to the total number, *i.e.* overall correct prediction rate. It is calculated as follows

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

2.3.2. Recall

Recall is used to measure the ability of the model to correctly identify the positive class (the considered class), which is also known as true positive rate (TPR) and sensitivity in binary case. And it is defined as

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

For imbalanced data classification, recall is a convenience measure to evaluate the prediction accuracy on a specific considered class (minority or majority).

2.3.3. Precision

For all compounds predicted as certain considered mechanism group, precision is the percentage of phenols correctly identified. So it is calculated as

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

2.3.4. F-measure

When the performance of learning models on only a specific class is considered, *F*-measure is adopted as a synthetic measure. It is a more flexible index, which is given in Eq. (8), using a tunable parameter β to indicate the relative importance of recall and precision. The parameter β can be modified to place more emphasis on either recall or precision. Typically, $\beta = 1$ was used in the paper, which means that we assign equivalent importance to recall and precision. A high *F*-measure value ensures that both recall and precision are reasonably high. Hence, a better model has a higher *F*-measure value.

$$F\text{-measure} = \frac{(1 + \beta^2) \times \text{recall} \times \text{precision}}{\beta^2 \times \text{recall} + \text{precision}} \tag{8}$$

2.3.5. G-mean

When the balanced performance of learning models on a whole data set is considered, *G*-mean is adopted as a synthetic measure. The basic *G*-mean measure is designed for binary classification problems. It is a geometric mean of two recall values of both classes, in which both recalls are expected to be high simultaneously. Here, we considered a four-class problem for toxicity assessment. For a *k*-class scenario, motivated by the reference [25], we defined *G*-mean as

$$G\text{-mean} = \left(\prod_{i=1}^{k} \text{recall}_i \right)^{1/k} \tag{9}$$

where recall_i denoted the recall of class *i*.

2.3.6. AUC

The Receiver Operating Characteristic curve (ROC), is a plot to evaluating the quality of the predicted models, with the sensitivity on the *y*-coordinate *versus* 1-specificity on the *x*-coordinate. The true positive rates and the false positive rates would vary along with the decision threshold changing from 0 to 1. The Area Under Curve (AUC) of ROC curve was calculated. The better QSAR model is with the lager AUC. If the AUC is 1, a perfect classifier can be found. And the AUC equals 0.5, when the classifier is the same as random guess.

3. Results

All experiments were implemented by using the software package Weka, which is a useful machine learning toolkit [24].

The three methods were compared: the original Random Forest (RF in short), sampling-based cost-sensitive Random Forest (RF-S in short), and threshloding-based cost-sensitive Random Forest (RF-T in short). Each experiment was performed twice. At first, the *group* 1 was used for building a model, and the *group* 2 for model testing. Later, the reverse process was performed. Then the two sets of results were simply numeric averaged as the reported results. The cost matrix used for each experiment was adjusted by hand during training process, and the appropriate cost matrix with acceptable results was chosen.

3.1. Selection of descriptors

In this work, correlation-based feature selection with Best First search was applied to select optimal subset of molecular descriptors [26]. Six descriptors (E_{LUMO}, E_{HOMO}, pK_a, N_{Hdon}, number of NO_2 groups, polar surface area) were selected to build the models.

This selected subset was agreed with the previous researches. During selecting molecular descriptors, it was found that E_{LUMO}, E_{HOMO} and pK_a afford the main contribution to make toxicity mechanism prediction. Many previous QSAR investigations had pointed out the large impact of such descriptors on successfully modeling toxicity. Aptula et al. reported that E_{LUMO}, pK_a, and N_{Hdon} are most important descriptors for discrimination between modes of action of phenols based on individual molecular descriptors [11].

Reactive compounds typically tend to have smaller E_{LUMO} values than less reactive compounds. As an electrophilicity descriptor, E_{LUMO} is frequently used in toxicological modeling to quantify the reactivity of compounds. Further, pro-electrophiles require metabolic activation, which in case of oxidative pathways could be modeled by E_{HOMO}. And N_{Hdon} represents the capability of denoting a hydrogen-bond. As we know, polar narcotics are different from nonpolar narcotics because of the diversity of their ability of hydrogen-bond interactions. So, N_{Hdon} can be used as an important descriptor to search for polar narcotics.

The selection procedure in this study suggested us with the two additional descriptors: polar surface area and number of NO_2 groups. Polar surface area was a helpful attribute to discriminate polar narcotics from others, since polar narcotics tended to have much lower polar surface area values. Additionally, soft electrophiles had relatively larger number of NO_2 groups than polar

Table 4
Comparison on overall prediction accuracy.[a]

	RF-S	RF-T	RF	Aptula [11]	Ren [13]	Norinder-tree [16]	Norinder-ensemble [16]
Global	0.923	0.914	0.859	0.873	0.852	0.718	0.927
Local							
1	0.945	0.950	0.905		0.922	0.782	0.914
2	0.968	0.959	0.950	0.939	0.794	0.973	0.968
3	0.964	0.950	0.909	0.946	0.788	0.918	0.982
4	0.968	0.959	0.959		0.711	0.950	0.968

[a] 1, polar narcotics; 2, weak acid respiratory uncouplers; 3, precursors to soft electrophiles; 4, soft electrophiles.

Table 5
Comparison on AUC.

	RF-S	RF-T	RF
Local			
Polar narcotics	0.991	0.962	0.931
Respiratory uncouplers	0.991	0.928	0.877
Pro-electrophiles	0.989	0.957	0.742
Soft electrophiles	0.988	0.957	0.954

Table 7
Comparison on *G*-mean.

	RF-S	RF-T	RF
Global	0.885	0.902	0.701
Local			
Polar narcotics	0.949	0.951	0.885
Respiratory uncouplers	0.957	0.926	0.774
Pro-electrophiles	0.931	0.955	0.707
Soft electrophiles	0.963	0.958	0.853

narcotics and pro-electrophiles. Hence, number of NO_2 groups was an effective descriptor to distinguish soft elecrophiles from others.

3.2. Performance report

A global model for all the four mechanisms of toxic action and four local models for a specific mechanism were constructed and evaluated. The performance report of all the prediction models was shown in Tables 4, 5, 6 and 7.

The original Random Forest model achieved appreciable overall accuracy with the range from 85.9% to 95.9% (Table 4). This result was competitive with discriminant analysis models whose overall prediction accuracies varied between 83.3% and 88.7% [11], decision tree models in the study of Ren with an averaged global accuracy 85.2% [13] and in the study of Norinder et al. with 71.8% accuracy [16]. These studies adopted the same data set as that used in our work. Particularly, Aptula et al. and Norinder et al. used the two-fold cross-validation with the group partition like that used in this work. The correct prediction rates of the Random Forest local models are also comparable to those previous studies (Table 4).

The original Random Forest model failed to solve the imbalanced data problem. It performed poorly on the minority classes, although it achieved an appreciable overall accuracy. The conclusion can be obtained from the analysis on recall. The recall measures the proportion of correctly classified phenols in a specific class. It was a way to measure the discriminative ability of the model on a specific class. In global Random Forest model, the recalls for the three rare classes were much lower than that of the majority class (polar narcotics) (Table 6). Additionally, on the whole data set, the *G*-mean of Random Forest model was also lower than those of cost-sensitive models (Table 7). For local models, the AUC value of polar narcotics was much higher than those of the rare class local models (Table 5). AUC was a more sophisticated tool to measure the discriminative ability of the model on a specific class.

The overall discriminative abilities of prediction models were improved after the two cost-sensitive strategies were applied. The overall accuracies of the cost-sensitive models were better than those of the original Random Forest models (Table 4). The *G*-mean and AUC values of the cost-sensitive models were much higher (Tables 5 and 7). Particularly, the cost-sensitive local model with sampling mode for respiratory uncouplers achieved the best AUC value 0.99.

The discriminative ability on a specific rare class was greatly improved by cost-sensitive Random Forest. This can be

Table 6
Comparison on recall, precision and *F*-measure.[a]

	Recall			Precision			*F*-measure		
	RF-S	RF-T	RF	RF-S	RF-T	RF	RF-S	RF-T	RF
Global									
1	0.941	0.914	0.961	0.979	0.993	0.885	0.960	0.948	0.921
2	0.778	0.833	0.611	0.824	0.833	0.688	0.800	0.830	0.648
3	0.963	1.000	0.556	0.765	0.730	0.833	0.848	0.844	0.670
4	0.870	0.870	0.739	0.870	0.800	0.810	0.870	0.834	0.773
Local									
1									
P	0.956	0.956	0.838	0.878	0.890	0.851	0.915	0.922	0.844
N	0.941	0.947	0.934	0.979	0.980	0.928	0.960	0.963	0.931
2									
P	0.970	0.965	0.980	0.995	0.990	0.966	0.982	0.977	0.973
N	0.944	0.889	0.611	0.739	0.696	0.733	0.829	0.780	0.667
3									
P	0.974	0.948	0.964	0.984	0.995	0.935	0.979	0.971	0.949
N	0.889	0.963	0.519	0.828	0.722	0.667	0.857	0.825	0.583
4									
P	0.970	0.959	0.984	0.995	0.995	0.970	0.979	0.977	0.959
N	0.957	0.957	0.739	0.786	0.733	0.850	0.857	0.823	0.701

[a] 1, polar narcotics; 2, weak acid respiratory uncouplers; 3, precursors to soft electrophiles; 4, soft electrophiles; P, positive class; N, negative class.

Table 8
Confusion matrices of RF-S, RF-F and RF with different cost weights.

	RF-S				RF-T								RF	
	(a)		(b)		(c)		(d)		(e)		(f)		(g)	
Cost weight	1:3		1:12		1:2		1:3		1:4		1:5		1:1	
Confusion matrix	184	9	186	7	184	9	187	6	183	10	179	14	186	7
	5	22	2	25	4	23	2	25	0	27	0	27	13	14

summarized from recall, precision, *F*-measure and *G*-mean results (Tables 6 and 7). For the global model, the recalls of the three rare classes were promoted. It was the similar case for the local models. Particularly, for the local model of pro-electrophiles, the thresholding-based cost-sensitive learning promotes the recall value from 51.9% to 96.3%. All the precision measures of cost-sensitive methods were much better than those of cost-insensitive methods. It indicated that the identified compound set for a specific mechanism was more pure. It was the similar case for *F*-measure and *G*-mean as recall and precision. It was worthy of highlighting the fact that the fourth local cost-sensitive model had a lower precision for the majority class, but it had a higher *F*-measure than that of the original model. It was confirmed that the cost-sensitive models keep better balance between recall and precision. It was in conformity with its higher AUC and overall accuracy.

Our results were competitive with the previous work. This can be found from the overall prediction accuracy results (Table 4) and the recall results. Aptula's global recall was 95% for polar narcotics, 70% for respiratory uncouplers, 75% for pro-electrophiles and 81% for soft electrophiles [11]. Norinder's global ensemble model recall was 95% for polar narcotics, 78% for respiratory uncouplers, 93% for pro-electrophiles and 87% for soft electrophiles [16].

However, there were negative results. The performance improvement on minority classes was along with the decrease of discriminative ability on the majority class in some cases. The recall values of cost-sensitive global Random Forest models for polar narcotics were lower than that of the original Random Forest. This fact was related to the cost matrix selection for the two types of misclassifications.

Additional prediction models with different cost matrices were constructed to investigate the relationship between the two types of misclassifications. The local models of pro-electrophiles were considered only, where pro-electrophiles were viewed as positive class. The cases for other three mechanisms were similar. The original, sampling-based and thresholding-based versions were built. In a cost matrix, $C(0,0)$ and $C(1,1)$ are both often viewed as zeros. So the effective factor in a cost matrix is the ratio of $C(1,0)$ to $C(0,1)$, which was called as cost weight in this study. The original Random Forest did not use a cost weight; it can be viewed as that Random Forest assigned equal weights to the two types of mistakes. The original Random Forest misclassified nearly one half of pro-electrophiles (13 out of 27), while the prediction accuracy of other mechanisms group is high to 96.4% (Table 8). The sampling-based model assigned more penalization to the second mistake, and found out much more true pro-electrophiles. The thresholding-based model had a similar result. With the second penalization $C(0,1)$ increasing, the prediction accuracy of pro-electrophiles became better. It was needed to note the fact that the performance on the majority class deteriorated. More majority class compounds were misclassified along with a relatively less weight assigned to them. In thresholding-based models with cost weights 1:4 and 1:5, all pro-electrophiles were found out. However, the performance on the majority class continued to deteriorate with a less weight 1:5. It was concluded that the prediction accuracies varied with the change of cost weights. What is the appropriate cost weight to be chosen in practical application? It would be discussed in the next section.

4. Discussion

Our findings confirmed that Random Forest is a useful tool for phenol toxicity mechanisms prediction with an appreciable accuracy. Random Forest had a competitive overall performance with other commonly used methods for biological data sets. The overall performance result (Table 4) extended the range of possible tools for phenol toxicity mechanisms prediction. However, the original Random Forest had failed to deal with the imbalanced data problem, like other common methods in machine learning field.

Cost-sensitive learning framework helped Random Forest promote the ability of the QSAR model to handle the imbalanced data problem. This finding strengthened our confidence on applying machine learning methods to biochemical data, which often have the imbalance problem. The discriminative ability of the QSAR model is a basic factor to apply *in silico* prediction methods to practical biochemical data exploring. If the QSAR model has a high prediction accuracy on major classes, but performs poorly on rare classes, this type of QSAR model has little practical sense. From the experimental results (Tables 5, 6 and 7), cost-sensitive Random Forest models had a better performance on rare classes than that of original Random Forest model. This finding suggested a possible promising way to handle imbalance problem in practical toxicity mechanisms prediction.

However, we must consider the fact that the performance on the majority class deteriorated a little in some cases. With the global model to be considered, the recalls of RF-S and RF-T were both a little lower than that of original Random Forest. This result indicated directly that more polar narcotics (the majority class) were misclassified as rare class compounds. The similar case can be found in Table 8. The reason of this phenomenon was that we had assigned more weights to the second type of mistakes. Compounds were misclassified as pro-electrophiles with more losses than other three classes. So compounds had more opportunities to be identified as pro-electrophiles.

The key practical question was which cost weight should be adopted. Without the loss of generality, pro-electrophiles were supposed as our interested objects. Then the practical requirement was to identify actual pro-electrophiles as many as possible, with some false pro-electrophiles to be tolerated. Here, for thresholding-based models, 1:4 was a good choice. If the practical goal was to keep an appropriate balance between these two types of misclassifications, the cost weight 1:3 should be adopted. Due to the limited space, the case of the global model for all 4-mechanism prediction was not analysed here. However, the case is similar to the local models. Above all, the cost weight was determined by domain experts according to the application goal. Exploring adaptive optimal cost weight searching methods will be the possible future work.

5. Conclusion

Our goal of this study was to develop mechanisms of toxic action prediction models on the TPT data set with a heavily skewed class distribution. One global and four local classification models were constructed by employing Random Forest in the cost-sensitive

learning framework. The optimal subset selection of molecular descriptors was involved, and the selected descriptors were agreed with the previous research conclusions. The statistical results in the paper confirmed that Random Forest was a competitive tool for building classification models of toxicity mechanisms prediction. However, the original Random Forest failed to handle the imbalanced data problem. Our investigation highlights the merits of using Random Forest learner in the cost-sensitive learning framework. The experimental results indicated the presented models indeed improve the performance on rare classes. However, the results also showed sometimes it brought a little degraded performance on the majority class. Fortunately, the balance of these two types of misclassifications can be adjusted by using cost matrices. It is anticipated that the cost-sensitive Random Forest learner may be promising for applications in toxicity assessment and other QSAR fields.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (60873092 and 90820306), the Natural Science Foundation of Chongqing (CSTC, 2009AB5002 and CSTC, 2010BB2217), and the Fundamental Research Funds for the Central Universities (Project No. CDJXS10182216).

References

[1] C.A. Rice-Evans, L. Packer, Flavonoids in Health and Disease, Marcel Dekker Inc., New York, 1998.
[2] S. Kar, A.P. Harding, K. Roy, P.L.A. Popelier, QSAR with quantum topological molecular similarity indices: toxicity of aromatic aldehydes to *Tetrahymena pyriformis*, SAR QSAR Environ. Res. 21 (2010) 149–168.
[3] H. Zhu, A. Tropsha, D. Fourches, A. Varnek, E. Papa, P. Gramatica, et al., Combinatorial QSAR modeling of chemical toxicants tested against *Tetrahymena pyriformis*, J. Chem. Inf. Model. 48 (2008) 766–784.
[4] D. Garg, T. Gandhi, C.G. Mohan, Exploring QSTR and toxicophore of hERG K^+ channel blockers using gfa and hypogen techniques, J. Mol. Graph. Model. 26 (2008) 966–976.
[5] P.C. Nair, M.E. Sobhia, Comparative QSTR studies for predicting mutagenicity of nitro compounds, J. Mol. Graph. Model. 26 (2008) 916–934.
[6] L. He, P.C. Jurs, Assessing the reliability of a QSAR model's predictions, J. Mol. Graph. Model. 23 (2005) 503–523.
[7] N. Stojić, S. Erić, I. Kuzmanovski, Prediction of toxicity and data exploratory analysis of estrogen-active endocrine disruptors using counter-propagation artificial neural networks, J. Mol. Graph. Model. 29 (2010) 450–460.
[8] H. Yuan, Y.Y. Wang, Y.Y. Cheng, Mode of action-based local QSAR modeling for the prediction of acute toxicity in the fathead minnow, J. Mol. Graph. Model. 26 (2007) 327–335.
[9] I. Kahn, S. Sild, U. Maran, Modeling the toxicity of chemicals to *Tetrahymena pyriformis* using heuristic multilinear regression and heuristic back-propagation neural networks, J. Chem. Inf. Model. 47 (2007) 2271–2279.
[10] S.J. Enoch, M. Hewitt, M.T.D. Cronin, S. Azam, J.C. Madden, Classification of chemicals according to mechanism of aquatic toxicity: an evaluation of the implementation of the Verhaar scheme in Toxtree, Chemosphere 73 (2008) 243–248.
[11] A.O. Aptula, T.I. Netzeva, I.V. Valkova, M.T.D. Cronin, T.W. Schultz, R. Kuhne, et al., Multivariate discrimination between modes of toxic action of phenols, Quant. Struct. Act. Rel. 21 (2002) 12–22.
[12] M. Jalali-Heravi, A. Kyani, Comparative structure–toxicity relationship study of substituted benzenes to *Tetrahymena pyriformis* using shuffling-adaptive neuro fuzzy inference system and artificial neural networks, Chemosphere 72 (2008) 733–740.
[13] S.J. Ren, Phenol mechanism of toxic action classification and prediction: a decision tree approach, Toxicol. Lett. 144 (2003) 313–323.
[14] B. Niu, Y.H. Jin, W.C. Lu, G.Z. Li, Predicting toxic action mechanisms of phenols using Adaboost learner, Chemometr. Intell. Lab. 96 (2009) 43–48.
[15] F.X. Cheng, J. Shen, Y. Yu, W.H. Li, G.X. Liu, P.W. Lee, et al., In silico prediction of *Tetrahymena pyriformis* toxicity for diverse industrial chemicals with substructure pattern recognition and machine learning methods, Chemosphere 82 (2011) 1636–1643.
[16] U. Norinder, P. Lidén, H. Boström, Discrimination between modes of toxic action of phenols using rule based methods, Mol. Divers. 10 (2006) 207–212.
[17] T.W. Schultz, G.D. Sinks, M.T.D. Cronin, Identification of Mechanisms of Toxic Action of Phenols to *Tetrahymena pyriformis* from Molecular Descriptors, SETAC Press, Pensacola, USA, 1997, pp. 329–342.
[18] M.T.D. Cronin, A.O. Aptula, J.C. Duffy, T.I. Netzeva, P.H. Rowe, I.V. Valkova, et al., Comparative assessment of methods to develop QSARs for the prediction of the toxicity of phenols to *Tetrahymena pyriformis*, Chemosphere 49 (2002) 1201–1221.
[19] C.L. Russom, S.P. Bradbury, S.J. Broderius, D.E. Hammermeister, R.A. Drummond, Predicting modes of toxic action from chemical structure: acute toxicity in the fathead minnow (pimephales promelas), Environ. Toxicol. Chem. 16 (1997) 948–967.
[20] Y. Xue, H. Li, C.Y. Ung, C.W. Yap, Y.Z. Chen, Classification of a diverse set of *Tetrahymena pyriformis* toxicity chemical compounds from molecular descriptors by statistical learning methods, Chem. Res. Toxicol. 19 (2006) 1030–1039.
[21] T. Olsson, V. Sherbukhin, SELMA, Synthesis and Structure Administration (SaSA), AstraZeneca R&D, Mölndal, Sweden, 1997.
[22] L. Breiman, Random forest, Mach. Learn. 45 (2001) 5–32.
[23] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, second ed., Wadsworth, Pacific Grove, CA, 1984.
[24] I.H. Witten, E. Frank, M.A. Hall, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, third ed., Morgan Kaufmann Publishers, 2010.
[25] R.P. Espíndola, N.F.F. Ebecken, On extending F-measure and G-mean metrics to multi-class problems, Data Min. VI 35 (2005) 25–34.
[26] M.A. Hall, Correlation-based feature selection for discrete and numeric class machine learning, in: International Conference on Machine Learning, Morgan Kaufmann Publishers, Stanford University, CA, 2000, pp. 359–366.