Three Data Mining Techniques To Improve Lazy Structure−Activity Relationships for Noncongeneric Compounds

Selina Sommer

Lehr- und Forschungseinheit für Bioinformatik, Ludwig-Maximilians-Universität München, Amalienstrasse 17, D-80333 München, Germany

Stefan Kramer*

Institut für Informatik I12, Technische Universität München, Boltzmannstrasse 3, D-85748 Garching b. München, Germany

Received January 19, 2007

We present three simple, yet effective data mining techniques for lazy structure−activity relationships (SARs) of noncongeneric compounds. In lazy SARs, classifications are particularly tailored for each test compound. Therefore, it is possible to make the most of the structure of a test compound. In our case, we derive its substructures and use them to determine similar structures. To obtain a well-balanced and representative set of structural descriptors, we enrich this set by strongly activating or deactivating fragments from the training set and subsequently remove redundant fragments. Finally, we perform *k*-Nearest Neighbor classification for several values of *k* and take a vote among the resulting predictions. These techniques (enrichment, removing redundancy, and voting) are integrated into the system iSAR (instance-based structure−activity relationships) and tested individually to show the relative contribution to the system's performance. Experiments on three data sets indicate that this simple and lightweight approach performs at least on the same level as other, more complex approaches.

1. INTRODUCTION

In structure−activity relationships for noncongeneric compounds we generally have to cope with the lack of knowledge about modes of actions or cellular targets. Therefore, the task of creating a model for the prediction of activity is essentially a data mining problem.[1] The focus of this paper is on structure−activity relationships for toxicological and environmental endpoints, with unknown or only partially known targets, and databases of heterogeneous structures. In principle, we can follow two strategies for the classification of new compounds: First, we build a model based on a training set and then apply it to all unseen cases to make predictions. This strategy is called *eager learning* in the machine learning literature.[2] Second, and alternatively, we can consider each test instance individually and extract information from the training set specifically for the prediction of that instance. Since the learning step is delayed to the testing phase, this strategy is usually called *lazy learning*. One common lazy learning scheme is instance-based learning, in particular *k*-Nearest Neighbor classification, where a test instance is classified according to the most similar instances from the training set. The main advantage of lazy learning is that it is possible to make the most of the information about a test instance.

In particular for structure−activity relationships based on substructures or molecular fragments, lazy learning should be helpful: Since the number of conceivable substructures or fragments is vast, it should be useful to restrict them to only those occurring in a test compound. Lazy structure−activity relationships have been shown to perform well recently by Helma.[3] In this paper, we show how lazy structure−activity relationships for noncongeneric compounds can be improved by three simple data mining techniques. The goal of the improvements is to obtain a well-balanced, nonredundant set of descriptors, and a robust classification, insensitive to variations in parameter settings. In the first step, we choose a feature set particularly tailored for a test instance to be classified from the instance itself as well as from the training set. Next, we remove redundancy (with respect to the data) from the feature set. Finally, we improve the classifications by making *k*-Nearest Neighbor less sensitive to the choice of a particular value *k*.

In our experiments, we show the effect of each of the three techniques on three data sets, DSSTox/CPDB mutagenicity, biodegradability, and the NCI DTP HIV data. Next, we show how the integrated approach, iSAR (instance-based structure−activity relationships), performs compared to established methods and results from the literature. Most importantly, we compare the approach to lazar, the lazy SAR method recently proposed by Christoph Helma.

This paper is organized as follows: In section 2, we present the materials, in particular the data sets used in our study, and the methods, from feature generation via feature selection to classification. In section 3, the approach is evaluated experimentally, before we draw our conclusions in section 4.

2. MATERIALS AND METHODS

2.1. Data. We use three different data sets of noncongeneric compounds for testing our methodology. This section

10.1021/ci600560m CCC: $37.00
Published on Web 10/06/2007

Table 1. Characteristics of the Data Sets Used

	CPDB(739[a])		biodegradability (322[a])		HIV (1481[a])	
classes	mutagenic	nonmutagenic	resistant	degradable	confirmed active	confirmed moderately active
number of compounds	355	384	140	182	410	1071
av number of atoms	13.6	15.1	12.1	10.0	39.9	31.8
min. number of atoms	2	2	2	2	10	6
max. number of atoms	64	90	26	29	189	222
av number of bonds	13.8	15.2	12.6	9.9	42.6	34.3
min. number of bonds	1	1	1	1	10	5
max. number of bonds	65	96	31	31	196	234
mean distances nearest neighbor	0.0564		0.0542		0.0849	
mean distances nearest hit	0.0704		0.0681		0.1036	
mean distances nearest miss	0.1140		0.1082		0.1676	

[a] Number of compounds.

introduces the data sets, and Table 1 presents some overview statistics concerning the size and composition of the data sets and compounds. For each data set, we are given a two-class problem: each compound is assigned to one of the two classes "active" or "inactive", where activity is defined according to the given endpoint.

2.1.1. Carcinogenic Potency Database (CPDB). The CPDB[4−6] is a database integrated into the Distributed Structure-Searchable Toxicity (DSSTox) public database network, which is a project of EPA's computational toxicology program. The version we use (CPDBAS_v2a_1451_1Mar05) contains structures of 1451 chemicals, which are classified for rodent carcinogenicity and mutagenicity according to the Salmonella assay. We use the mutagenicity data for classification. Those compounds evaluated as mutagenic or weakly mutagenic in ref 7 or as overall positive in ref 8 are defined as mutagenic in this database, all others evaluated for this endpoint as nonmutagenic. After removing duplicates and all compounds without mutagenicity information, 739 elements remain, of which 355 are classified as mutagenic (active) and 384 as nonmutagenic (inactive) with respect to Salmonella mutagenicity.

2.1.2. Biodegradability Data Set (biodeg). This data set contains rates of biodegradation in terms of half-life time for 342 chemicals, compiled by Howard et al.[9] The authors derived the degradation rates for different types of degradation either from measurements or from expert estimations, if experimental data were not available.

The biodegradability data set was studied in depth by Blockeel et al.,[10] where several SAR approaches from machine learning and inductive logic programming (ILP) were tested. The task is to predict biodegradation of chemicals in aqueous environment under aerobe conditions. In the data set, the measured or estimated half-life times as well as discretizations are given: The compounds are classified as "resistant" if the half-life time is longer than 4 weeks (defined as active) and "degradable" if the half-life time is shorter (defined as inactive). [Note that 28 days are required in some of the common tests for biodegradation, e.g., according to the OECD guidelines.] The data set contains 322 molecules after removing duplicates, 140 resistant and 182 degradable.

2.1.3. NCI DTP HIV Data Set (HIV). In the context of the AIDS Antiviral Screen of the NCI Developmental Therapeutics Program (DTP),[11] more than 40 000 chemicals were tested for anti-HIV activity. More precisely, a soluble formazan assay was used to measure protection of human CEM cells against HIV-1 infection (for details see the paper by Weislow et al.[12]). According to the outcomes of the screening, the compounds of the data set (October 1999 release) were divided into three classes: "confirmed active" (CA), "confirmed moderately active" (CM), and "confirmed inactive" (CI). A compound showing at least 50% or 100% protection in repeated tests is classified as CM or CA, respectively. All other tested molecules belong to the class CI. Obviously, several two-class problems can be derived from the data. Similar to related approaches in the literature,[13,14] we chose the task of distinguishing between confirmed active and moderately active compounds. This gives us 1481 chemicals without duplicates, 410 of class CA and 1071 of class CM.

2.1.4. Composition of Data Sets. The degree of difficulty of a data set for an instance-based SAR method depends to some extent on its composition, in particular, on whether it contains singletons, analog series, or mixtures thereof. Visual inspection of the data sets shows that the biodegradability and mutagenicity data contain mostly singletons. However, the mean similarity of structures is still quite high, which may be partly due to the relatively small size of the structures. The HIV data set contains many more analogs than the other two data sets. To quantify this, we calculate the mean distance (over all compounds of a data set) to the nearest neighbor according to our distance measure. Moreover, we compute the mean distance to the nearest hit and to the nearest miss. The nearest hit of an instance is the nearest instance of the same class, whereas the nearest miss is the nearest instance of a different class. The intuition behind these measures is that in data sets with analogs, the mean distance to the nearest neighbor should be quite small. If classification is easy, the difference between the mean distance to the nearest hit and the mean distance to the nearest miss should be large on average. As can be seen in the last three rows of Table 1, the values are more or less the same for CPDB mutagenicity and biodegradability. However, quite contrary to our expectations, the mean distance to the nearest neighbor is higher in the HIV data. Also, the difference between the latter two measures is larger as well but not much larger if normalized by the mean distance to the nearest neighbor. This suggests that the HIV data set is not easier or harder than the first two. In fact, looking at the results (section 3), the improvement over the baseline accuracy is only roughly 9%, compared to values greater than 20% for

the CPDB and biodegradability. Therefore, we may conclude that the HIV endoint is not easier to predict, even though the presence of analog series should make it easier for instance-based learning methods. This fact is also illustrated by an example structure shown at the end of section 3.

2.2. General Approach. The concept of structure−activity relationships is founded on the key assumption that the activity of a molecule depends on its structural properties. Therefore, one possible approach is to use substructures (molecular fragments) as descriptors for the prediction of activity. We apply *lazy learning* techniques with a feature set specifically modified for each test instance, i.e., each molecule to be classified. Lazy learning, as opposed to *eager learning*, omits the step of building a classification model in advance, that is, during training. Instead, lazy learning algorithms delay model creation to the classification phase or even do not create any model at all.[15] We use the latter type of lazy learning, which is called instance-based learning. The features used for describing the molecules are the occurrences of substructures, in our case, two-dimensional linear fragments or in graph-theoretical terms, *paths*. [Paths were used successfully in a number of approaches[1,3] and systematically compared to trees and general subgraphs recently.[16,17] As information about targets or mechanisms is lacking in many toxicological and environmental applications, more precise 3D representations or force field approaches often do not improve performance. This may be due to overfitting, given the limited data, a large number of energetically possible conformations,[18] and the uncertainty in the endpoint.]

Our approach of predicting the activity of molecules using instance-based classification combined with instance-based feature selection is divided into three completely decoupled steps: the generation of a basic feature set, the selection of the features appropriate for a given test instance, and finally the classification of this test instance. Before we go into detail, we briefly present the data mining techniques on a general level.

Given a structural representation of molecules, it is clear that a vast number of descriptors could be generated, with only a few of them relevant for the classification of a compound at hand. The first and most straightforward way of reducing the number of possible descriptors is considering only those fragments present in the test compound. [This principle was already applied in the CASE/MultiCASE system.[19,20]] However, this information may not be enough to determine useful distances from the compounds of the training set. Since the absence of highly activating or deactivating fragments can also give hints about the similarity to other compounds and thus the activity, we enrich the set of fragments occurring in the test instance "in small doses" by molecular fragments predominantly present in active (or inactive) compounds of the training set.

However, the resulting set of structural fragments may still contain redundant information. If one fragment is contained in another, and the former is part of exactly the same compounds as the latter, then the former should be considered as redundant, i.e., it does not provide additional structural information with respect to a given set of compounds. We therefore propose to remove such redundancy in order to improve the performance in terms of runtime as well as predictive accuracy.

Finally, we circumvent the choice for a particular k in k-Nearest Neighbor classification by taking a vote across several parameter values. In other words, several values for k are tested (as specified by the user), and the classifications are combined using a simple voting procedure. This corresponds to the idea of ensembles, as known to be effective in machine learning and data mining. In the following, we present the three techniques employed by iSAR in more detail.

2.3. Feature Generation. In the following, we assume a 2D graph representation of molecular structure and consider subgraphs, in our case paths, as substructural features. In the feature generation step, we compute all subgraphs with a frequency of occurrence greater than a user-defined threshold. The frequency is counted per example: Multiple occurrences within one molecule are only counted once. Given the full data set, we create a basic feature set, from which an appropriate subset is selected for the classification of each test instance.

The paths are generated using the graph mining algorithm gSpan' by Katharina Jahn,[21] an optimization of the gSpan algorithm[22] for molecular graphs. Although being developed for mining subgraphs, it is also possible to restrict the output to trees and paths.

The minimum frequency constraint for the mining process is chosen very low in order to not exclude probably important features in advance. For the CPDB and the biodeg data, all paths occurring in at least two molecules of the data set are generated. In the case of the HIV data, we set the minimum support to an absolute value of 5, taking into account the size and the complexity of the compounds in this data set.

2.4. Feature Selection. Feature selection is one of the most important steps in the prediction process. If the features relevant for class membership are outnumbered by many irrelevant ones in a high-dimensional feature space, the prediction will become error-prone. If, on the other hand, the selected feature set is too small, it might not provide enough information to decide about the target class.

An important aspect of the prediction process is that the feature selection step is completely decoupled from the feature generation. This is done for efficiency: The computation of substructures frequent in the training set has to be done only once and provides the basis for the classification of all test instances. For a given test instance, the relevant substructures are selected from this pool as needed. The feature selection starts from the whole set of paths generated in the previous step, using the criteria of importance described in the following. Figure 1 shows the data flow of selecting features for a test instance schematically.

As in some text mining approaches (e.g., for spam detection[23]), features are selected instance-based, so that for each test instance a different feature space is used. At first, those paths are extracted that occur in both the current test compound and the basic feature set resulting from the feature generation step. However, this can leave for some test instances no or only few features, especially for very small structures.

As also the absence of a highly activating or deactivating substructures of a molecule can give a hint about its classification, we incorporate the nonoccurrence of strongly discriminating substructures as additional features. The most important question for this part of the feature selection step

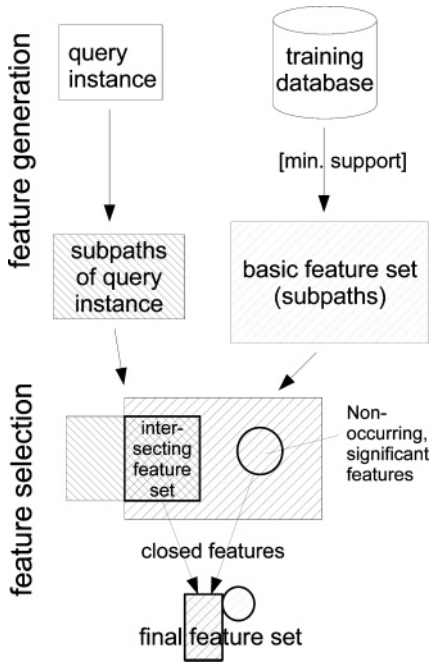

Figure 1. Data flow of feature generation and selection in iSAR: All paths occurring in a test instance and the basic feature set are selected and additionally highly significant substructures not occurring in the test instance. The final feature set consists of the closed features within this selected set.

is to find an adequate value balancing the number of occurring and nonoccurring features optimally. Having tested various settings for adding nonoccurrences, we chose the one performing best on average for all three data sets: adding the 0.5% most relevant features according to their statistical significance determined by the χ^2 test, or by the exact Fisher test, if the χ^2 test is not applicable.

The basic feature set usually contains many paths that occur in exactly the same subset of molecules of the data set. In a set of paths with the same occurrences, often one path may be part of another (i.e., it may be *more general*). This means that these sets contain much redundant information, which can harm classification performance. By selecting only *closed features*, the described redundancies are removed. A feature is called *closed*,[24] if it is not more general than any other feature occurring in the same examples. In the context of paths, a feature *A* is *more general* than a feature *B*, if the path corresponding to *A* is a subpath of the one of *B*. On the other hand, paths with the same occurrences, but without the generality relationship, are kept because their co-occurrence may be incidental and the information may be nonredundant.

2.5. Classification. For the classification step, we use the Euclidean distance metric, and each vote of the *k* neighbors of a test instance x_{test} is weighted by similarity ($w(n) = 1 - d(n, x_{test})$). As the optimal value for the parameter *k* changes depending on the context, it does not make much sense to pick one particular value of *k*. A common way of choosing an appropriate *k* is to perform an internal round of cross-validation on the training set. Unfortunately, this procedure would not work well for our instance-based feature selection: As the feature set is specifically adapted to the current test instance, evaluating the performance of different *k* on the training instances would be misleading.

We therefore adopted the idea from ensemble theory not to rely on a single classification but to combine several classifications, each from a different value for *k*. In this way, the classification can be made more independent of the choice of a particular value of *k*. It has been noted in several papers, e.g., by Pfahringer,[25] that ensembles of classifiers often outperform single classifiers.

Let *K* be the set of different values for *k* specified by the user, for instance, $K = \{3, 4, 5, 10, 20\}$. Moreover, we assume that the classes are numbered from 1 to *l*. Now a classifier returns, for a given test instance x_{test}, a vector $\hat{\mathbf{y}} = (\hat{y}_1, ..., \hat{y}_l)$, with the class probability estimates $\hat{y}_i$ for each class *i* ($\sum_{i=1}^{l} \hat{y}_i = 1$). Since we have $|K|$ different *k*-NN classifiers, we have as many class probability estimates. The class probability estimates for a particular value of *k* are denoted by $\hat{\mathbf{y}}^k$. From these class probability estimates, a joint prediction is made by simple averaging over the individual class probability estimates.

More precisely, for a single *k*-NN classifier with a given value of *k*, the prediction of a class *i* is made by

$$\hat{y}_i^k = P_k(y = i) = \frac{\sum_{n \in \{\text{neighbors} | y = i\}} w(n)}{\sum_{n \in \{\text{neighbors}\}} w(n)} \quad (1)$$

In other words, $\hat{y}_i^k$ is the estimated probability that class *y* takes value *i* according to the nearest neighbor classifier with *k* neighbors.

As mentioned above, the overall class is calculated from a set of $|K|$ such classifiers, where the combined classification, the classification *c*, is determined as follows:

$$c = \text{argmax}_{i \in \{1,...,l\}} \sum_{k \in K} \hat{y}_i^k \quad (2)$$

The quality of this combined classification depends on the set of classifiers contributing to it. Like for the classification of a single classifier, also for the combined classification, a probability of correctness can be derived. The probability for a class $\hat{y}_i$ of the combined prediction can be estimated by calculating the average probability of this class according to the set of classifiers:

$$\hat{y}_i = P_{\text{combined}}(y = i) = \sum_{k \in K} P_k(y = i)/|K| = \sum_{k \in K} \hat{y}_i^k/|K| \quad (3)$$

The pseudocode in the algorithm of Chart 1 formalizes the complete prediction procedure, from feature generation to classification.

3. RESULTS

3.1. Feature Sets and Runtime Performance. In Table 2, we present statistics of the feature generation and selection procedure. In the upper section of the table, the size of the data sets and the minimum frequency parameters for substructure search in absolute and relative terms can be found. The parameters are set as small as possible, not to miss any rare fragment that may be relevant for classification. The middle section of the table contains statistics on the

Chart 1. Algorithm 1: iSAR Prediction Procedure

procedure iSAR($x_{test}, trainSet, minSup, K$)

// ***feature generation***
// call $gSpan'$ to compute all paths with a
// minimum frequency $minSup$ in the training
// dataset
$basicFeatureSet \leftarrow$
gSpan'($trainSet, minSup$)
// ***feature selection***
// select features occurring in test instance
$F \leftarrow \{f \in basicFeatureSet | f$ occurs in $x_{test}\}$
// enrich feature set by significant
// non-occurring features
$F \leftarrow F \cup \{f \in basicFeatureSet | f$ is
significantly associated with the
class in $trainSet$ according to
the χ^2 test$\}$
// remove redundant (non-closed) features
$F \leftarrow \{f \in F | is\_closed(f)\}$
// ***classification***
for all $k \in K$ **do**
// get class probability estimates for all
// classes from kNN with parameter k
$\hat{\mathbf{y}}^k \leftarrow$ kNN($x_{test}, trainSet, k, F$)
end for
// return class with the maximal sum of
// estimated class probabilities across
// several values of k
return $argmax_{i \in \{1,\ldots,l\}} \sum_{k \in K} \hat{y}_i{}^k$

end procedure

Table 2. Numbers of Features and Running Times of Various Stages of iSAR

	CPDB	biodeg	HIV
instances	739	322	1481
minSup (absolute)	2	2	5
minSup (relative)	0.27%	0.62%	0.34%
total no. of features	4895	1988	62 349
no. of features/instance			
mean	49.1	39.1	581.3
std dev	77.0	48.4	947.4
no. of closed features/instance			
mean	33.1	19.6	197.7
std dev	32.5	16.7	138.0
final no. of features/instance			
mean	83.6	63.0	284.9
std dev	15.5	5.9	88.4
feature generation			
total time (s)	1	<1	115
feature selection			
total time (s)	2	1	41
time/instance (s)	0.001	0.003	0.028
classification			
total time (s)	213	30	3358
time/instance (s)	0.29	0.09	2.27

numbers of features passing each stage of the procedure. The lower section presents the running times in total and per instance.

Substructures of sufficient frequency are computed once before classification takes place and stored in an index, such that the substructures of an instance can be found quickly, and conversely, the instances containing a given substructure. Theoretically, the computation of all frequent subgraphs in a graph database is the most expensive step of the whole procedure, as it heavily builds on (in the general case NP-hard) subgraph isomorphism tests. However, practically the running times are acceptable. The running times of graph mining algorithms critically depend on the minimum frequency parameter, the size of the data set, and the average size and diversity of the graphs in the database. As can be seen in Table 2, the running times and numbers of resulting patterns for HIV are much higher than for the other two data sets, which can be explained by the analog series and the on average larger structures.

Once the substructural features are available, the computations per instance are theoretically and practically simple. The time it takes to compile the final feature set for an instance depends largely on the initial number of features. Subsequently, the data set is scanned once for all values of k simultaneously to determine the list of nearest neighbors. In other words, this step is, as in many other instance-based learning schemes, linear in the number of structures and linear in the size of the final feature set. This is also reflected in the practical running times of the classification per instance shown in the last row of Table 2.

3.2. Evaluation. In order to evaluate our classification results, we use leave-one-out (LOO) cross-validation, where each instance of the data set serves once in turn as test instance, while the remaining instances form the training set. Then, the results of the single classifications are aggregated to form the average accuracy or area under the receiver operating characteristic (ROC) curve. A ROC curve plots the true positive rate of a binary classifier against its false positive rate. The area under the ROC curve (AUC) is a measure of how well a classifier sorts positive and negative examples relative to each other. In our case of instance-based feature selection and classification, LOO cross-validation is a natural choice because each instance has to be classified separately, anyway. This means that applying LOO cross-validation does not increase the effort of classification as compared to, e.g., 10-fold cross-validation. The parameters values for k were set to $K = \{3, 4, 5, 10, 20\}$.

Table 3. Leave-One-Out Results of iSAR on the Three Data Sets (CPDB, Biodegradability, and HIV) and Results from 10-Fold Cross-Validation

	CPDB	biodeg	HIV
Leave-One-Out Results of iSAR			
accuracy	0.751	0.773	0.813
TP rate	0.749	0.807	0.715
TN rate	0.753	0.747	0.851
AUC	0.821	0.851	0.842
no. of instances	739	322	1.481
10-Fold Cross-Validation Results			
accuracy	0.745	0.770	0.812
TP rate	0.760	0.807	0.708
TN rate	0.730	0.741	0.852
AUC	0.820	0.848	0.840
no. of instances	739	322	1.481

For each of the three used data sets, we present the obtained accuracy and the true positive (TP) and true negative (TN) rate as well as the AUC and the number of classified instances in the upper half of Table 3. In the lower half, we also state the results from 10-fold cross-validation, to show that leave-one-out does not give overly optimistic estimates. In fact, the results from leave-one-out and 10-fold cross-validation are remarkably close. The ROC curves are shown in Figure 2.

Our methodology of combining closed path features with nonoccurrences performs very well for all three data sets. On average, it outperforms alternative parameter settings for the prediction process, which has been confirmed by comparative tests.[17] In the following, we evaluate the benefit

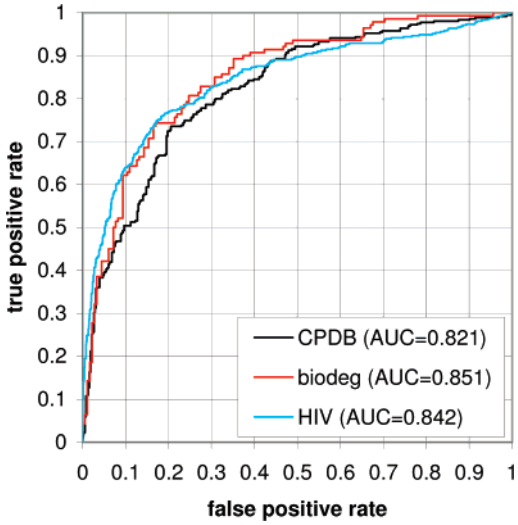

Figure 2. ROC curves for the three data sets CPDB, biodeg, and HIV.

Table 4. Comparison of Results for Weighted Combined Classification (iSAR) and *k*-NN Classification for a Single Exemplary Parameter Value $k = 5$

	CPDB		biodeg		HIV	
	iSAR	$k = 5$	iSAR	$k = 5$	iSAR	$k = 5$
acc.	0.751	0.731	0.773	0.767	0.813	0.809
AUC	0.821	0.804	0.851	0.837	0.842	0.837

Table 5. Comparison of Results for Using Only Closed Features (iSAR) and Features without This Restriction

	CPDB		biodeg		HIV	
	iSAR	also nonclosed	iSAR	also nonclosed	iSAR	also nonclosed
acc.	0.751	0.723	0.773	0.792	0.813	0.799
AUC	0.821	0.809	0.851	0.870	0.842	0.835

of the three main features of our approach: (a) combining several *k*-NN classifications, (b) adding highly significant nonoccurring features, and (c) using closed features only. We compare the classification results obtained when omitting one of these techniques to the original results presented above, each one in turn.

3.2.1. Benefit of Combined Classification. As stated above, the choice of an appropriate value for parameter *k* in *k*-Nearest Neighbor classification is critical to its success. In experiments not shown in detail in this paper, we found that a combined classification from various values of *k* outperforms *k*-NN with the best individual value of *k*.[17] To illustrate the effect, we picked one of the best overall values from our experiments, and set $k = 5$. Table 4 shows the results for the three data sets in terms of predictive accuracy and AUC.

The difference for both settings is highest for the CPDB data set, with an accuracy greater by two percentage points in favor of iSAR, and smallest for the HIV data set. Generally speaking, the weighted combined classification of iSAR meets or even exceeds the results of the best individual contributing *k*-NN classifier.[17]

3.2.2. Benefit of Using Only Closed Features. Next, we analyze the benefit of restricting the feature set to closed features. In Table 5, the iSAR results are compared to those obtained when using also nonclosed features, while leaving all other settings unchanged.

Table 6. Comparison of Results Using Nonoccurrences (iSAR) and Omitting Nonoccurrences as Features

	CPDB		biodeg		HIV	
	iSAR	no nonocc	iSAR	no nonocc	iSAR	no nonocc
acc.	0.751	0.729	0.773	0.755	0.813	0.796
AUC	0.821	0.786	0.851	0.806	0.842	0.843

The results are not consistent for all three data sets: While omitting the restriction to closed features increases the accuracy as well as the area under the ROC curve for the CPDB and the HIV data set, the values decrease in the case of the biodegradability data set. A very likely cause for the negative effect of using only closed features in the latter case is the very low number of nonredundant substructures occurring in a molecule of the biodeg data set on average (20 features). For instances with extremely few occurring features, the majority of features might be nonoccurring, making this information too unspecific for classification in some cases. The molecules of the other two data sets, however, contain on average more nonredundant features, and the positive effect of removing redundant information prevails. Apart from improving the results for two of the three data sets, using only closed features strongly reduces the number of features and with it the computational effort of classification.

3.2.3. Benefit of Adding Nonoccurring Features. Finally, we evaluate the impact of adding highly discriminative substructures not occurring in the test instance as features for classification. As presented in Table 6, using nonoccurring features additionally clearly improves the results for the CPDB and the biodeg data set, while the differences in the case of the HIV data set are only marginal.

The molecules of the HIV data set are very large and complex on average (see Table 1), which implies that they already contain many substructures, and relatively little information can be added by nonoccurrences. For the smaller and less complex compounds of the two other data sets, on the other hand, highly significant nonoccurring substructures can provide useful additional class information.

3.3. Comparison to Related Approaches. In this section, we compare the results of our technique to those of related approaches, in order to get an idea of how well it performs in the overall context of SAR prediction.

3.3.1. Lazar. Being also a lazy SAR approach, lazar[3] represents the closest work in the literaure. Lazar also uses instance-based feature selection and a *k*-NN related classification. Moreover, the availability of lazar allows for evaluation of it on the same compilation of data sets, which makes it convenient for a comparison. For the NCI DTP HIV data set, however, we are not able to obtain results from lazar because of the size and the complexity of the data causing main memory problems. Table 7 lists the results achieved by lazar together with those of our approach for the remaining two data sets, CPDB and biodeg. Figure 3 displays the corresponding ROC curves. Results are given for all classified instances and for those within the applicability domain (AD), which is determined based on the threshold of 0.05 confidence as defined by Helma.[3] In our case, the estimated probability of correctness of the clas-

Table 7. Comparison of Our Results with Lazar Results for the CPDB and the Biodegradability Data Set and Statistics for All Classifications and for the Applicability Domain (AD)

	iSAR		lazar	
CPDB	all	AD	all	AD
accuracy	0.751	0.891	0.755	0.831
TP rate	0.749	0.871	0.769	0.847
TN rate	0.753	0.910	0.742	0.811
AUC	0.821	0.893	0.805	0.849
no. of instances	739	266	706	266

	iSAR		lazar	
biodeg	all	AD	all	AD
accuracy	0.773	0.859	0.730	0.820
TP rate	0.807	0.882	0.706	0.866
TN rate	0.747	0.839	0.750	0.770
AUC	0.851	0.908	0.813	0.870
no. of instances	322	156	270	156

Table 8. (a) Comparison of Our Results with the Best Results of Blockeel et al. for the Biodegradability Data Set and (b) Comparison of Our Results with Those of Deshpande et al., Horvath et al., and Frasconi et al. for the NCI DTP HIV Data Set

(a)

biodeg	iSAR	C4.5 (P2)	log.reg. (P2)	C4.5 (P1+P2)	log.reg. (P1+P2)
acc.	0.773	0.722	0.784	0.762	0.748

(b)

HIV	iSAR	FSG[26,27]	CPK[14]	WDK[28]
AUC	0.842	0.810	0.827	0.842

sification determines if an instance belongs to the applicability domain (see also eq 3). In order to make the results comparable, we adapt the threshold for the applicability domain of iSAR, so that it classifies exactly the same number of instances as lazar.

In the case of the CPDB, lazar and our approach achieve very similar accuracies, with the value for lazar being slightly higher. The area under the ROC curve, however, is larger by two percentage points for iSAR. For the biodeg data set, we obtain significantly improved results regarding accuracy as well as ROC curves, which is remarkable in contrast to the relatively smaller differences observed for the CPDB data. Considering instances within the applicability domain, our results also clearly exceed those of lazar. Note that lazar does not classify 4% of the CPDB instances and 16% of the biodeg instances at all (i.e., even outside the applicability domain), while our approach returns classifications for the whole data set.

3.3.2. Biodegradability. Now, we compare our results for the biodegradability data set to those presented by Blockeel et al.[10] The paper tests several classification and regression methods on various representations of the data. For classification, accuracies and ROC curves are reported but not the areas under the curves. The descriptors include global features (molecular weight and logP), predefined functional groups (a feature set called P1 in the paper), and small automatically generated substructures (P2). From the 40 experiments in the paper, only one experiment shows superior performance compared to iSAR: logistic regression applied to the global features, P1 and P2. In Table 8(a), we present some of the best results, logistic regression and C4.5 applied

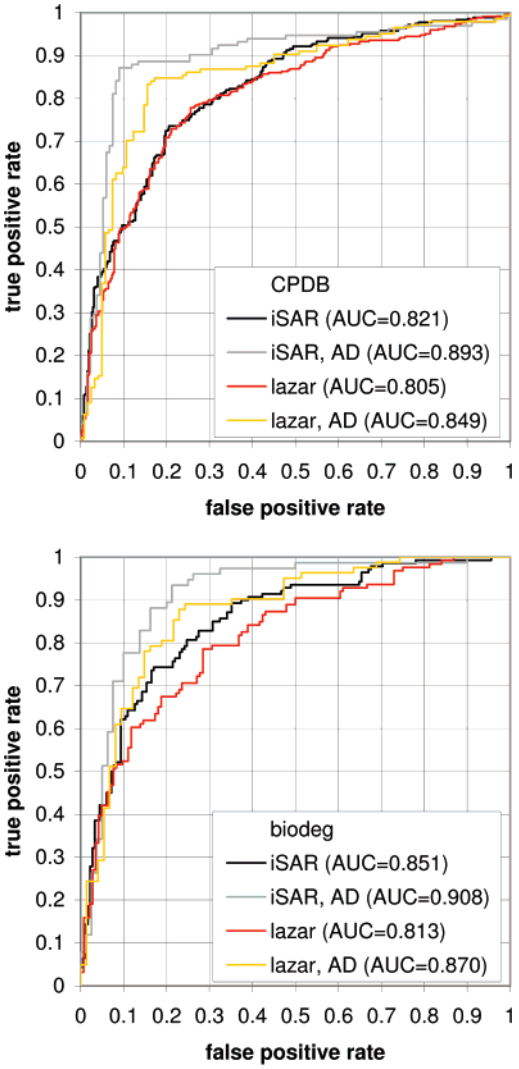

Figure 3. Comparison of ROC curves for lazar and iSAR.

to P1 and P1+P2. These accuracies show that iSAR achieves a strong performance on the data set, without having to test dozens of combinations of machine learning approaches and sets of descriptors.

3.3.3. Support Vector Machines. In order to compare our approach to eager learning techniques, we apply the Support Vector Machine (SVM) implementation of the WEKA workbench to the same data sets. However, in these tests, SVMs show a predictive performance substantially inferior to our instance-based approach.[17] This finding implies that it is necessary to adapt algorithms for the classification of molecular or generally structured data. This has been successfully done for SVMs by developing kernel functions specialized for structured datalike graphs, for instance by Horvath et al.[14] and Frasconi et al.[28] A drawback of the kernel-based approaches, however, is their high computational burden. Moreover, the prediction of structure−activity relationships using SVMs tends to receive a low acceptance in some application areas, because of their black-box nature, and thus problems with explaining classifications. In the field of toxicology, for instance, users require an explanation for classifications in risk assessment. [iSAR is able to provide both relevant substructures and relevant instances for each of its predictions. For kernel methods for structured data, it is in principle possible to return the support vectors and

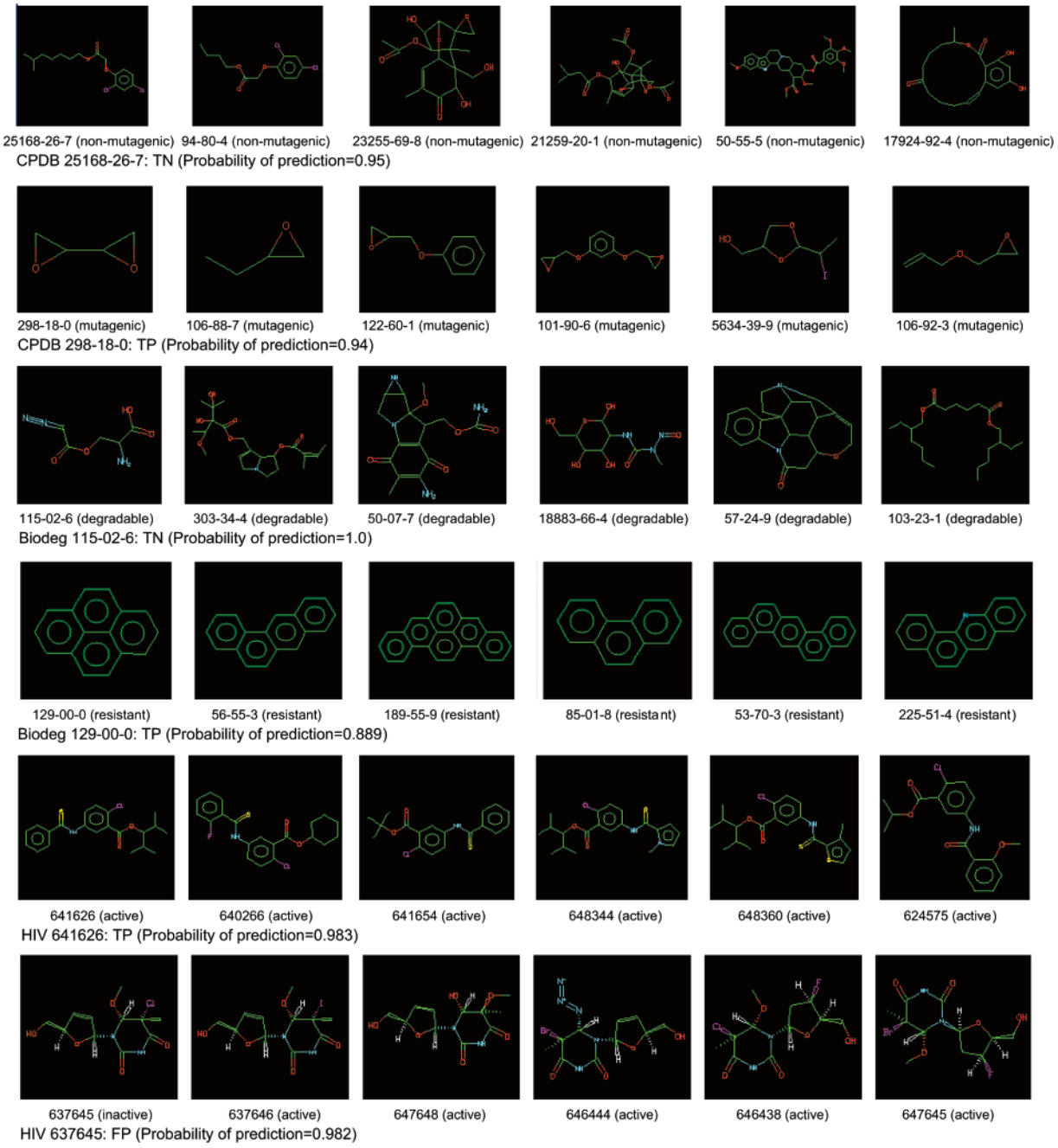

Figure 4. Two examples of iSAR classifications for each data set.

kernel evaluations; however, this output is somewhat less intuitive.]

The NCI DTP HIV data set has been used for evaluation in several publications. Deshpande et al.[26,27] evaluate their SAR approach based on this data set. Moreover, we include the results for the above two kernel-based approaches, weighted-decomposition kernels by Frasconi et al. and cyclic pattern kernels by Horvath et al. Both approaches are evaluated using support vector machines for different classification problems. For a detailed description, we refer to the original publications.

The results given as AUC values in Table 8 show that iSAR substantially outperforms the approach presented in Deshpande et al.[26,27] Moreover, it proves to be comparable to much more complicated and computationally demanding techniques based on kernels.

3.4. Sample Predictions. In the following, we briefly present a few sample predictions made by iSAR. To do so, we randomly sampled ten compounds from each of the three data sets. From those ten compounds, we picked the two compounds with the highest predicted probability for a class. In Figure 4, the test compounds are shown along with their five nearest neighbors. Following the test compound in the left-most column, the five nearest neighbors are listed in ascending order of the distance. Visual inspection shows that most of the suggested similarities can be interpreted from a chemical point of view, in many cases also with respect to the endpoints of, e.g., mutagenicity and biodegradability.

In all cases except test compound one and three (25168-26-7 and 115-02-6), similarities in terms of larger structural units can be recognized, although the measure is essentially fingerprint-based.[29] However, 25168-26-7 and 115-02-6

mainly feature, *local similarities*, that is, cases where smaller fragments can be mapped onto each other.

It should be noted that the three nearest neighbors have the greatest impact on the classification in our setting ($K = \{3, 4, 5, 10, 20\}$), as these data points are used five times, whereas more remote points are used just once (neighbors ranked 11−20) or twice (neighbors ranked 6−10). Therefore a suitable similarity to the first three neighbors is the most critical factor of success.

Finally, the example in the last row shows how misleading the analogs in the HIV data set can be. Although all of the nearest neighbors with a clear similarity suggest an active compound, the actual classification is "moderately active". This shows that distinguishing actives and moderately actives is a hard task, because they differ only in the reproducibility of anti-HIV activity.

4. CONCLUSION

We presented a new lazy-learning approach to structure−activity relationships of noncongeneric compounds. The targeted applications are SARs for toxicological and environmental endpoints, with databases of heterogeneous structures. Three simple techniques from data mining are used to obtain a well-balanced set of substructure features particularly tailored for a test instance to be classified, to remove redundant substructures, and to obtain robust classifications, without committing oneself to a single value of k in k-Nearest Neighbor learning. These techniques are integrated into a new system for instance-based structure−activity relationships called iSAR. In the paper, we focused on a quantitative evaluation of the approach and tested each technique individually to determine the relative contribution to the overall performance. The experiments show that, although being extremely simple, these techniques are effective in practice. In fact, the results are as good as or better than the best results known from the literature, at moderate computational costs. As the computational burden is small compared to most other approaches, it should be particularly suited for very large data sets of noncongeneric compounds.

Although the three techniques were presented in the context of iSAR, they should be useful in isolation as well (depending on the application). In future work, we are planning to improve the representation of chemical information. The iSAR package will be made available on the Web as open source software.

ACKNOWLEDGMENT

We would like to thank Christoph Helma for providing the lazar package for an experimental comparison. Moreover, we thank Bernhard Pfahringer, Jörg Wegner, and Lothar Richter for valuable comments on an earlier draft of the paper.

REFERENCES AND NOTES

(1) Helma, C.; Cramer, T.; Kramer, S.; De Raedt, L. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1402-1411.
(2) Mitchell, T. *Machine Learning*; McGraw-Hill: 1997.
(3) Helma, C. *Mol. Diversity* **2006**, *10*, 147-158.
(4) Berkeley Lab. *The Carcinogenic Potency Project*; 2005.
(5) Gold, L. S.; Slone, T. H.; Ames, B. N.; Manley, N. B.; Garfinkel, G. B.; Rohrbach, L. Carcinogenic Potency Database. In *Handbook of Carcinogenic Potency and Genotoxicity Databases;* Gold, L. S., Zeiger, E., Eds.; CRC Press: Boca Raton, FL, 1997; Chapter 1, pp 1−605.
(6) Gold, L. S.; Manley, N. B.; Slone, T. H.; Rohrbach, L. *Environ. Health Perspect.* **1999**, *107* (Suppl. 4), 527−600.
(7) Zeiger, E. Genotoxicity Database. In *Handbook of Carcinogenic Potency and Genotoxicity Databases*; Gold, L. S., Zeiger, E., Eds.; CRC Press: Boca Raton, FL, 1997; Chapter 5, pp 687-729.
(8) Kier, L. E.; Brusick, D. J.; Auletta, A. E.; Von Halle, E. S.; Brown, M. M.; Simmon, V. F.; Dunkel, V.; McCann, J.; K. Mortelmans, M. P.; Rao, T. K.; Ray, V. *Mutat. Res.* **1986**, *168*, 69−240.
(9) Howard, P.; Boethling, R.; Jarvis, W.; Meylan, W.; Michalenko, E. *Handbook of Environmental Degradation Rates*; Lewis Publishers: 1991.
(10) Blockeel, H.; Dzeroski, S.; Kompare, B.; Kramer, S.; Pfahringer, B.; van Laer, W. *Appl. Artif. Intell.* **2004**, *18*, 157−181.
(11) National Cancer Institute. *DTP AIDS Antiviral Screen*; 2005.
(12) Weislow, O. S.; Kiser, R.; Fine, D. L.; Bader, J. P.; Shoemaker, R. H.; Boyd, M. R. *J. Natl. Cancer Inst.* **1989**, *81*, 577−586.
(13) Deshpande, M.; Kuramochi, M.; Wale, N.; Karypis, G. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 1036-1050.
(14) Horváth, T.; Gärtner, T.; Wrobel, S. Cyclic pattern kernels for predictive graph mining. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: 2004.
(15) Simoudis, E.; Aha, D. W. *Artif. Intell. Rev.* **1997**, *11*.
(16) Bringmann, B.; Zimmermann, A.; De Raedt, L.; Nijssen, S. Don't Be Afraid of Simpler Patterns. In *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*; Fürnkranz, J., Scheffer, T., Spiliopoulou, M., Eds.; Springer: 2006; Vol. 4213.
(17) Sommer, S. Lazy Structure-Activity Relationships, 2006 Diplomarbeit, Ludwig-Maximilians-Universität and Technische Universität München.
(18) Cronin, M. T. Toxicological Information for Use in Predictive Modeling: Quality, Sources, and Databases. In *Predictive Toxicology*; Helma, C., Ed.; Marcel Dekker: 2005.
(19) Klopman, G. *J. Am. Chem. Soc.* **1984**, *106*, 7315−7321.
(20) Klopman, G. *Quant. Struct.-Act. Relat.* **1992**, *11*, 176−184.
(21) Jahn, K. Graph Mining for Bio- and Cheminformatics, 2005 Diplomarbeit, Ludwig-Maximilians-Universität and Technische Universität München.
(22) Yan, X.; Han, J. gSpan: Graph-based substructure pattern mining. In *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM)*; IEEE Computer Society: 2002.
(23) Pfahringer, B. A Semi-Supervised Spam Mail Detector. In *Proceedings of the ECML/PKDD 2006 Discovery Challenge Workshop*; Bickel, S., Ed.; Humboldt University: Berlin, 2006.
(24) Pasquier, N.; Bastide, Y.; Taouil, R.; Lakhal, L. Discovering Frequent Closed Itemsets for Association Rules. In *Proceedings of the 7th International Conference on Database Theory (ICDT 1999)*; Springer: 1999.
(25) Pfahringer, B. (The futility of) trying to predict carcinogenicity of chemical compounds. In *Proceedings of the Predictive Toxicology Challenge Workshop, Twelfth European Conference on Machine Learning (ECML)*; 2001.
(26) Deshpande, M.; Kuramochi, M.; Karypis, G. *Frequent sub-structure-based approaches for classifying chemical compounds*; Technical Report 03-016; University of Minnesota: 2003.
(27) Deshpande, M.; Kuramochi, M.; Karypis, G. Frequent Sub-Structure-Based Approaches for Classifying Chemical Compounds. In *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM)*; IEEE Computer Society: 2003.
(28) Menchetti, S.; Costa, F.; Frasconi, P. Weighted Decomposition Kernels. In *Proceedings of the International Conference on Machine Learning (ICML)*; 2005.
(29) Raymond, J.; Willett, P. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 59−71.

CI600560M