Applied Artificial Intelligence, 18:157–181, 2004

ISSN: 0883-9514 print/1087-6545 online
DOI: 10.1080/08839510490279131

Taylor & Francis
Taylor & Francis Group

EXPERIMENTS IN PREDICTING BIODEGRADABILITY

HENDRIK BLOCKEEL
Department of Computer Science, Katholieke Universiteit Leuven, Leuven, Belgium

SAŠO DŽEROSKI
Department of Intelligent Systems, Jozef Stefan Institute, Ljubljana, Slovenia

BORIS KOMPARE
Faculty of Civil Engineering and Geodesy, University of Ljubljana, Ljubljana, Slovenia

STEFAN KRAMER*
Department of Computer Science, Technische Universität München, München, Germany

BERNHARD PFAHRINGER
Department of Computer Science, University of Waikato, Hamilton, New Zealand

WIM VAN LAER
Department of Computer Science, Katholieke Universiteit Leuven, Leuven, Belgium

This paper is concerned with the use of AI techniques in ecology. More specifically, we present a novel application of inductive logic programming (ILP) in the area of quantitative structure-activity relationships (QSARs). The activity we want to predict is the biodegradability of chemical compounds in water. In particular, the target variable is the half-life for aerobic aqueous biodegradation. Structural descriptions of chemicals in terms of atoms and

*The work described in this paper was conducted while the author was at the University of Freiburg, Machine Learning Lab. Georges-Köhler-Allee Geb. 079, D-79110 Freiburg i. Br., Germany.

Hendrik Blockeel is a post-doctoral fellow of the Fund for Scientific Research of Flanders. This work was supported in part by the ESPRIT IV Project 20237 ILP2. Thanks are due to Irena Cvitanič for help with preparing the data set in computer-readable form, Christoph Helma for help in preparing the background knowledge and calculating logP, and Ross King and Ashwin Srinivasan for providing some of the definitions of the functional group predicates and for providing some feedback on this work.

Address correspondence to Hendrik Blockeel, Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan 200A, B-3001 Leuven, Belgium. E-mail: Hendrik.Blockeel@cs.Kuleuven.ac.be

bonds are derived from the chemicals' SMILES encodings. The definition of substructures is used as background knowledge. Predicting biodegradability is essentially a regression problem, but we also consider a discretized version of the target variable. We thus employ a number of relational classification and regression methods on the relational representation and compare these to propositional methods applied to different propositionalizations of the problem. We also experiment with a prediction technique that consists of merging upper and lower bound predictions into one prediction. Some conclusions are drawn concerning the applicability of machine learning systems and the merging technique in this domain and the evaluation of hypotheses.

The persistence of chemicals in the environment (or to environmental influences) is welcome only until the time the chemicals fulfill their role. After that time, or if they happen to be in the wrong place, the chemicals are considered pollutants. In this phase of their life span, we wish that the chemicals would disappear as soon as possible. The most ecologically acceptable (and a very cost-effective) way of *disappearing* is the degradation of components that are not considered pollutants (e.g., mineralization of organic compounds). Degradation in the environment can take several forms, from physical pathways (erosion, photolysis, etc.), through chemical pathways (hydrolysis, oxydation, diverse chemolises, etc.) to biological pathways (biolysis). Usually the pathways are combined and interrelated, thus making degradation even more complex. In our study, we focus on biodegradation in an aqueous environment under aerobic conditions, which affects the quality of surface and ground water.

The problem of properly assessing the time needed for ultimate biodegradation can be simplified to the problem of determining the half-life time of that process. However, few measured data exist and often these data are not taken under controlled conditions. It follows that an objective and comprehensive database on biolysis half-life times can not be found easily. The best we were able to find was in a handbook of degradation rates (Howard et al. 1991). The chemicals described in this handbook were used as the basis of our study.

Usually, authors try to construct a QSAR (quantitative structure-activity relationship) model/formula for only one class of chemicals, or congeners of one chemical, e.g., phenols. This approach to QSAR model construction has an implicit advantage that only the variation with respect to the class mainstream should be identified and properly modeled. Contrary to the described situation, our database comprises several families of chemicals, e.g., alcohols, phenols, pesticides, chlorinated aliphatic and aromatic hydrocarbons, acids, ketones, ethers, other diverse aromatic compounds, etc. From this point of view, the construction of adequate QSAR models/formulae is a much more difficult task.

We apply several machine learning methods, including several inductive logic programming methods, to the above database in order to construct

SAR/QSAR models for biodegradability. This application is discussed both from the biochemical and the machine learning viewpoint.

GOALS OF THIS PAPER

From the biochemical point of view, the main point of this article is to illustrate the applicability of machine learning in general and inductive logic programming in particular in the context of biodegradability.

From the machine learning point of view, this paper is a case study in which we consider several machine learning methods and approaches in the specific context of biodegradability prediction. We classify machine learning methods along several dimensions and study the effect of these dimensions on the performance of systems in this domain.

More specifically, we are looking for an answer to the following questions.

- How does the use of different representations for the data influence the performance of machine learning systems?
- Prediction problems like this one are essentially numerical, but can be stated as a classification problem. To what extent is it advantageous to use a direct regression approach instead of an indirect classification approach (as defined and explained below)?
- How do different machine learning methods (rule set induction, decision tree induction, and statistical approaches) compare?

Concerning the *data representation*, the main issue we want to investigate is to what extent the greater representational power of inductive logic programming is advantageous in this domain. We distinguish three different kinds of representation:

- A propositional representation, where each molecule is described by stating some properties of the molecule as a whole that domain experts expect to be relevant.
- A representation where molecules are described by a fixed list of attributes, and these attributes themselves are generated automatically using some kind of feature construction (which may be of a trivial nature). In this paper, we will refer to this representation as the *propositionalized* representation because the feature construction is essentially obtained by generating certain kinds of queries that an ILP system would typically generate and storing the results of these queries as attributes of the examples. Note that when following this approach, the representation issue is decoupled from the induction issue. ILP is used to generate a good propositional representation, but from then on only propositional techniques are used.

An obvious question arising is whether this decoupling harms predictive performance, as compared with a full ILP approach.

- A relational representation, where molecules are described by listing all atoms and bonds in the molecule with their properties and the relationships between them (i.e., which atoms participate in which bonds). Some information about substructures (benzene rings, etc.) is also represented in this manner.

The second issue is that of using *regression versus classification* methods. A regression method directly predicts the target value (which is numerical), whereas a classification method predicts a class derived from the target value. Clearly, if the goal of the prediction is to accurately predict the half-life time itself, then classification is not of any use; however, in the context of biodegradability, it is not so important to know exactly how fast a chemical will degrade, but rather whether it will degrade within a reasonable time span. In this context, classification does make sense, while regression methods are still applicable as well. Thus the question arises: Is there any advantage in using regression methods instead of classification methods? Will the more precise information on half-life times that regression systems automatically use help them to provide better classification?

The third issue is to what extent the machine learning *paradigm* to which the system belongs matters. In this respect, we compare rule-based systems, tree-based systems, and systems that are directly based on statistics (linear regression, logistic regression, naïve Bayes).

DATA SET

The database used was derived from the data in the handbook of degradation rates (Howard et al. 1991). The authors have compiled the degradation rates for 342 widely used (commercial) chemicals from the available literature. Where no measured data on degradation rates were available, expert estimations were provided. The main source of data employed was the Syracuse Research Corporation's (SRC) Environmental Fate Data Base (EFDB), which in turn used as primary sources of information DATALOG, CHEMFATE, BIOLOG, and BIODEG files to search for pertinent data.

For each considered chemical, the book contains degradation rates in the form of a range of half-life times (low and high estimate) for overall, biotic, and abiotic degradation in four environmental compartments, i.e., soil, air, surface water, and ground water. We focus on surface water here. The overall degradation half-life is a combination of several (potentially) present pathways, e.g., surface water photolysis, photooxydation, hydrolysis, and biolysis (biodegradation). These can occur simultaneously and have even synergetic effects, resulting in a half-life time (HLT) smaller than the HLT for each

of the basic pathways. We focus on biodegradation here, which was considered to run in unacclimated aqueous conditions, where biota (living organisms) are not adapted to the specific pollutant considered. For biodegradation, three environmental conditions were considered: aerobic, anaerobic, and removal in waste water treatment plants (WWTP). In our study, we focus on aqueous biodegradation HLT's in aerobic conditions.

The HLT's in the original database of Howard et al. (1991) are given in hours, days, weeks, and years. In our database, we represented them in hours. We took the arithmetic mean of the low and high estimate of the HLT for aqueous biodegradation in aerobic conditions: The natural logarithm of this mean was the target variable for machine learning systems that perform regression. In additional experiments, we have also used the natural logarithm of the upper and lower bounds themselves as target variable (see experimental section).

A discretized version of the arithmetic mean was also considered in order to enable us to apply classification systems to the problem. Originally (Džeroski et al. 1999) four classes were defined: chemicals degrade *fast* (mean estimate HLT is up to seven days), *moderately* fast (one to four weeks), *slowly* (one to six months), or are *resistant* (otherwise). In the experiments described here, we further abstract from these four classes and define a two-class problem. More precisely, a compound is considered to *degrade* if its class is *fast* or *moderate*; otherwise, it is considered *resistant*.

From this point on, we proceeded as follows. The CAS (Chemical Abstracts Service) registry number of each chemical was used to obtain the SMILES (Weininger 1988) notation for the chemical. In this fashion, the SMILES notations for 328 of the 342 chemicals were obtained.

The SMILES notation contains information on the two-dimensional structure of a chemical. So, an atom-bond representation, similar to the representation used in experiments to predict mutagenicity (Srinivasan et al. 1996), can be generated from a SMILES encoding of a chemical. A DCG-based translator that does this has been written by Michael De Groeve and is maintained by Bernhard Pfahringer. We used this translator to generate atom-bond relational representations for each of the 328 chemicals. Note that the atom-bond representation here is less powerful than the QUANTA-derived representation, which includes atom charges, atom types, and a richer selection of bond types. The types especially carry a lot of information on the substructures of which the respective atoms/bonds are a part.

A global feature of each chemical is its molecular weight. This was included in the data. Another global feature is logP, the logarithm of the compound's octanol/water partition coefficient, used also in the mutagenicity application. This feature is a measure of hydrophobicity, and can be expected to be important since we are considering biodegradation in water.

The basic atom and bond relations were then used to define a number of background predicates defining substructures/functional groups that are possibly relevant to the problem of predicting biodegradability. These predicates are: nitro ($-NO_2$), sulfo ($-SO_2$ or $-O-S-O_2$), methyl ($-CH_3$), methoxy ($-O-CH_3$), amine, aldehyde, ketone, ether, sulfide, alcohol, phenol, carboxylic_acid, ester, amide, imine, alkyl_halide (R-Halogen where R is not part of a reasonant ring), ar_halide (R-Halogen where R is part of a resonant ring), epoxy, n2n ($-N=N-$), c2n ($-C=N-$), benzene (resonant C_6 ring), hetero_ar_6_ring (resonant 6 ring containing at least 1 non-C atom), non_ar_6c_ring (non-resonant C_6 ring), non_ar_hetero_6_ring (non-resonant six ring containing at least one non-C atom), six_ring (any type of six ring), carbon_5_ar_ring (resonant C_5 ring), non_ar_5c_ring (non-resonant C_5 ring), non_ar_hetero_5_ring (non-resonant five ring containing at least one non-C atom), and five_ring (any type of five ring). Each of these predicates has three arguments: MoleculeID, MemberList (list of atoms that are part of the functional group), and ConnectedList (list of atoms connected to atoms in MemberList, but not in MemberList themselves).

EXPERIMENTS

Goals

We previously discussed the goals of this study; the experiments will, of course, reflect these. More specifically, our experiments are set up in order to enable a comparison between different *machine learning systems*, between different *problem representations*, and between *classification and regression*, as well as an assessment of the usefulness of machine learning methods in the domain of biodegradability.

The experimental setup should be such that results are maximally informative with respect to the above questions. We now describe this setup in more detail.

Representations

We distinguish four different constituents of the data representations, which we refer to as *Global, P1, P2*, and *R*.

- *Global* contains global descriptors of molecules that experts assume to be relevant. In our experiments, we used the molecular weight (mweight) and the logarithm of the octanol/water partition coefficient of the molecule (logP).
- *P1* contains counts of the substructures and functional groups listed at the end of the previous section.

- *P2* contains counts of automatically generated small substructures (all connected substructures of two or three atoms, and those of four atoms that have a star-topology).
- *R* contains a description of the whole molecular structure: atoms, bonds, and the substructures from *P1* (not only their counts, but more precisely described by listing the atoms occurring in them and the atoms through which they are attached to the rest of the molecule).

Note that *P1* and *P2* are human-defined propositionalizations, i.e., a human expert defined which substructures could be of interest, then these substructures were found in the compounds using a relatively simple algorithm. We have not experimented with discovery-based propositionalization methods such as Warmr (Dehaspe and Toivonen 1999), although this would be worthwhile to investigate in further work.

By considering all possible combinations of these chunks of background knowledge, a lattice of different representations is obtained (partially ordered by the *contains less information than* relation), as shown in Figure 1. Starting from *Global*, where no relational information is used at all, one can add chunks of relational information (or information derived from relational information) one by one, finally obtaining the most informative background $Global + P1 + P2 + R$.

Language Bias of Machine Learning Systems

Most ILP systems use a declarative language bias specification to decide how to make use of certain information. For propositional systems, this is much less the case, because in the attribute-value formalism, the way information is used is very much standardized (comparison of attributes with constants). Therefore, while the above lattice of background information

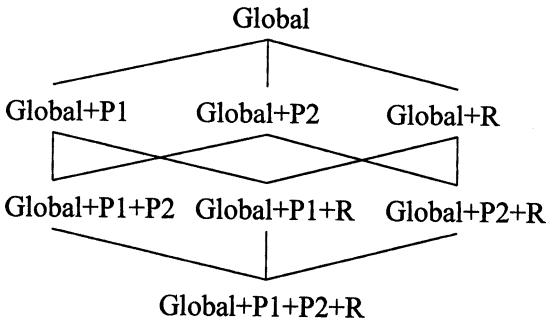

FIGURE 1 A lattice of background information.

is sufficient to guarantee that propositional systems will use the same information (and hence can be accurately compared in this respect), for ILP systems it is also necessary to describe their language bias, that is, exactly what information they use and how they use it.

The fact that ILP systems use different bias specification languages slightly complicates this: How does a bias specification for one system compare to that of another? We have decided to use the following approach: The language bias is specified as precisely as possible in a natural language, then the users of the different ILP systems write a language specification that conforms to this informal specification. This approach turned out to work quite well.[1]

The bias specifications used for the ILP systems are as follows:

- Global: Allow inequality comparisons of molecular weight or logP value with constants generated using the discretization procedure of the system. The number of discretization thresholds was chosen to be eight.
- P1: Allow *equality* and *greater than* tests for the number of times a specific substructure occurs in the compound. When in combination with R, make sure also to introduce the list of atoms through which the substructure is connected to the rest of the compound. (These atoms can possibly later be used in other tests).
- P2: Allow *equality* and *greater than* tests for the number of times a specific substructure occurs in the compound.
- R: Allow the following tests, in which *specific* means that a constant should be filled in here and *some* means that an existentially quantified variable is to be filled in:

- whether a specific element occurs in the compound;
- whether a given atom is of a specific element type;
- whether a specific bond occurs in the compound;
- whether a specific bond occurs between given atoms or between a given atom and any other atom;
- whether some bond occurs between given atoms;
- whether some or a specific bond between a given atom and a new atom of some specific element occurs; and
- whether the list of atoms connecting a given substructure to the rest of the compound contains a specific element.

Note that other types of test could be used by ILP systems as well, such as testing whether two substructures touch. The above list of tests was chosen based upon the certainty we had that a) the tests are meaningful to domain experts and b) they can be accurately specified in the language bias specification of the different systems.

Systems

A variety of classification and regression systems were applied to the classification and the regression version of the biodegradability problem. Table 1 sorts them according to whether they can handle relational data or not, whether they are tree-based, rule-based, or based more directly on statistics, and whether they can handle regression or classification.

Propositional systems were applied to all propositional representations (i.e., all combinations excluding R). For classification, these were the decision tree inducer C4.5 (Quinlan 1993b), its rule generating add-on C4.5rules (Quinlan 1993b), logistic regression, and a naive Bayesian classifier. For regression, linear regression was used as well as the regression-tree induction program M5′ (Wang and Witten 1997) and a reimplementation of M5′ (Quinlan 1993a). M5′ constructs linear models in the leaves of the tree.

Relational learning systems applied include ICL (De Raedt and Van Laer 1995), which induces classification rules, S-CART (Kramer 1996, 1999), and TILDE (Blockeel and De Raedt 1998). The latter are capable of inducing both classification and regression trees. ICL is an upgrade of CN2 (Clark and Boswell 1991) to first-order logic, TILDE is an upgrade of C4.5, and S-CART is an upgrade of CART (Breiman et al. 1984). TILDE cannot construct linear models in the leaves of its trees; S-CART can.

Regarding parameter settings, default settings were employed for all systems except for S-CART and Tilde where the stopping criterion of tree induction was adapted manually based on experience with earlier experiments (in the case of S-CART to generate larger trees; in the case of Tilde to generate smaller trees [F-test at 0.05]). Besides language bias, no parameter settings were varied throughout the experiments described here.

Design of Experiments

Two different induction tasks were considered in these experiments:

- Classification into *degradable* and *resistant*.
- Prediction of the mean HLT estimated by the experts.

TABLE 1 Systems Used in Experiments, Classified Along Three Dimensions

		Tree-based	Rule-based	Statistical
classification	prop.	C4.5	C4.5rules	log. regr., NB
	rel.	S-CART, Tilde	ICL	
regression	prop.	M5′		linear regression
	rel.	S-CART, Tilde		

The experiments were designed orthogonally with respect to systems, backgrounds, induction tasks, and train/test-partitionings. More specifically, for each system, a tenfold cross-validation was run for each different background in the lattice where it was applicable, for each induction task for which it was applicable, and on each of five different 10-fold partitionings of the data. This orthogonality provides maximal flexibility with respect to the statistical tests that can be performed (e.g., as the same train/test sets are used it is possible to perform paired comparisons between the systems).

Evaluation Criteria

We evaluated the predictive models that were induced according to the following criteria. For classification, accuracy (number of correct predictions divided by total number of predictions) was used. For regression systems, both the Pearson correlation coefficient (between predictions and actual values) and the root mean squared error (RMSE) were computed.

In order to be able to compare regression systems with classification systems, an ROC (Receiver Operating Characteristics) analysis (Provost and Fawcett 1998) was performed. This ROC analysis is one of the reasons why the classification task was stated as a two-class problem, instead of the four-class problem considered in earlier work (Džeroski et al. 1999). The two-class classification is also frequently found in the literature and, from the domain expert's point of view, it is equally useful.

Comparison Tests

Classifiers were compared using McNemar's test for changes: For each individual instance, the prediction of classifiers A and B is compared to the real class of the instance; the number of times A is better than B is counted and compared with the number of times B is better than A. Under the null hypothesis that both classifiers are equally good, both number should be approximately equal. Precise statistical tests are available to test whether a deviation from this situation is significant.

Regression systems were compared using the sign test: For each instance, an algorithm scored a point if its prediction was closer to the target than that of another algorithm. Again, under a null hypothesis of both systems being equally good, both systems should score approximately the same.

In the presence of so many tests, it is not uncommon to apply Bonferroni adjustment to the significance levels. We are not doing this here because Bonferroni adjustment only makes sense when the different tests that are performed are independent, which is not the case here. As Dietterich (1998) argues, statistical tests, when used as we do here, are always to be interpreted as somewhat heuristic indications of differences in performance

levels, and we could not see good arguments to apply Bonferroni adjustment in this context. We do use a significance level of 0.01 for all tests.

RESULTS OF EXPERIMENTS

We now describe the experiments we have performed. There are two batches of experiments. In the first batch, a straightforward approach to classification and regression was followed. In an attempt to improve the quality of the produced models, we have run a second batch of experiments, using a novel approach that combines predictions for lower and upper bounds into an overall numerical prediction.

First Batch: Classification and Regression

In the first batch of experiments, systems were trained directly from the target values that should be predicted, i.e., since we want to predict the class or HLT of compounds, those attributes are considered target values for the learners.

In Table 2, classification accuracies are given for different systems. The results are shown in a lattice, in order to make it easier to compare a) performance of a particular system with different kinds of background knowledge, b) the performance of different systems under the same background knowledge, and c) the influence of a particular chunk of background knowledge on the average performance of all systems. The same is done for regression systems: Correlation coefficients are shown in Table 3 and RMSEs are shown in Table 4.

In this section of the text we just mention some observations; a discussion of what they might mean follows later.

- Observation 1: Compared with the very restricted set of global attributes, any extension of the background knowledge (whether it is P1, P2, or R) yields a significant improvement in performance. After this initial boost, however, adding more information does not improve performance any further.
- Observation 2: Comparing the increase in performance that P1, P2, and R individually generate when added to Global reveals no significant differences between them (i.e., none of them significantly outperforms any other; they all outperform Global though).

Statistical tests were performed to compare the performance of different systems with the same background knowledge, and different sets of background knowledge for the same system. No significant results were consistently obtained.[2] The strongest result we obtained was that logistic

TABLE 2 Classification Accuracies for Different Systems and Different Background Knowledge (Mean and Standard Deviation Over 5 10-fold Cross-Validations)

Global	
System	Mean (Dev)
ICL	0.663 (0.008)
Tilde	0.666 (0.013)
S-CART	0.633 (0.009)
C4.5	0.605 (0.025)
C4.5rules	0.604 (0.021)
N.Bayes	0.655 (0.009)
Log.Reg.	0.648 (0.005)

Global + P1		Global + P2		Global + R	
System	Mean (Dev)	System	Mean (Dev)	System	Mean (Dev)
ICL	0.718 (0.019)	ICL	0.729 (0.015)	ICL	0.748 (0.009)
Tilde	0.709 (0.017)	Tilde	0.726 (0.015)	Tilde	0.736 (0.011)
S-CART	0.716 (0.012)	S-CART	0.722 (0.011)	S-CART	0.726 (0.013)
C4.5	0.750 (0.013)	C4.5	0.722 (0.016)		
C4.5rules	0.738 (0.016)	C4.5rules	0.739 (0.020)		
N.Bayes	0.720 (0.010)	N.Bayes	0.725 (0.004)		
Log.Reg.	0.752 (0.012)	Log.Reg.	0.784 (0.008)		

Global + P1 + P2		Global + P1 + R		Global + P2 + R	
System	Mean (Dev)	System	Mean (Dev)	System	Mean (Dev)
ICL	0.723 (0.018)	ICL	0.732 (0.006)	ICL	0.726 (0.020)
Tilde	0.723 (0.023)	Tilde	0.741 (0.013)	Tilde	0.729 (0.014)
S-CART	0.722 (0.004)	S-CART	0.719 (0.009)	S-CART	0.712 (0.017)
C4.5	0.762 (0.023)				
C4.5rules	0.730 (0.015)				
N.Bayes	0.730 (0.007)				
Log.Reg.	0.748 (0.025)				

Global + P1 + P2 + R	
System	Mean (Dev)
ICL	0.715 (0.020)
Tilde	0.729 (0.011)
S-CART	0.713 (0.023)

regression with background P2 almost consistently (four out of five partitionings) performs significantly better than several (not all) other systems. This is still a relatively weak conclusion, and the fact that a similar result is not obtained for $P1 + P2$ raises the suspicion that this result may be accidental.

To compare the regression and classification approaches, we have performed an ROC analysis (Provost and Fawcett 1998). In brief, ROC analysis

TABLE 3 Pearson Correlations for Regression
In roman: results of batch 1; italic: results of batch 2

Global	
System	Mean (Dev)
Tilde	0.487 (0.020)
	0.495 (*0.015*)
S-CART	0.476 (0.031)
	0.478 (*0.016*)
M5′	0.503 (0.012)
	0.502 (*0.014*)
Lin.reg.	0.436 (0.004)
	0.437 (*0.005*)

Global + P1		Global + P2		Global + R	
System	Mean (Dev)	System	Mean (Dev)	System	Mean (Dev)
Tilde	0.596 (0.029)	Tilde	0.615 (0.014)	Tilde	0.616 (0.021)
	0.612 (*0.022*)		*0.619* (*0.021*)		*0.635* (*0.018*)
S-CART	0.563 (0.010)	S-CART	0.595 (0.032)	S-CART	0.605 (0.023)
	0.581 (*0.015*)		*0.636* (*0.015*)		*0.659* (*0.019*)
M5′	0.579 (0.024)	M5′	0.646 (0.013)		
	0.592 (*0.013*)		*0.646* (*0.014*)		
Lin.reg.	0.592 (0.014)	Lin.Reg.	0.443 (0.026)		
	0.592 (*0.013*)		*0.455* (*0.022*)		

Global + P1 + P2		Global + P1 + R		Global + P2 + R	
System	Mean (Dev)	System	Mean (Dev)	System	Mean (Dev)
Tilde	0.603 (0.023)	Tilde	0.622 (0.022)	Tilde	0.594 (0.019)
	0.624 (*0.022*)		*0.646* (*0.017*)		*0.621* (*0.022*)
S-CART	0.593 (0.021)	S-CART	0.606 (0.015)	S-CART	0.599 (0.028)
	0.624 (*0.014*)		*0.630* (*0.013*)		*0.640* (*0.026*)
M5′	0.655 (0.014)				
	0.663 (*0.011*)				
Lin.reg.	0.563 (0.023)				
	0.575 (*0.024*)				

Global + P1 + P2 + R	
System	Mean (Dev)
Tilde	0.595 (0.020)
	0.618 (*0.022*)
S-CART	0.606 (0.032)
	0.631 (*0.026*)

distinguishes two types of errors: predicting a negative as positive and predicting a positive as negative. Classifiers are thus evaluated in two dimensions: FP reflects the false positive rate (proportion of negatives predicted

TABLE 4 Root Mean Squared Errors for Regression
In roman: results of batch 1; italic: results of batch 2

Global	
System	Mean (Dev)
Tilde	1.380 (0.022)
	1.370 (0.017)
S-CART	1.398 (0.032)
	1.388 (0.018)
M5′	1.355 (0.011)
	1.356 (0.013)
Lin.Reg.	1.412 (0.004)
	1.411 (0.004)

Global + P1		Global + P2		Global + R	
System	Mean (Dev)	System	Mean (Dev)	System	Mean (Dev)
Tilde	1.285 (0.041)	Tilde	1.283 (0.026)	Tilde	1.265 (0.033)
	1.260 (0.030)		*1.270 (0.034)*		*1.231 (0.025)*
S-CART	1.342 (0.013)	S-CART	1.315 (0.048)	S-CART	1.290 (0.038)
	1.313 (0.021)		*1.240 (0.022)*		*1.198 (0.034)*
M5′	1.294 (0.036)	M5′	1.204 (0.019)		
	1.272 (0.019)		*1.201 (0.020)*		
Lin.Reg.	1.276 (0.019)	Lin.Reg.	1.556 (0.053)		
	1.274 (0.017)		*1.530 (0.040)*		

Global + P1 + P2		Global + P1 + R		Global + P2 + R	
System	Mean (Dev)	System	Mean (Dev)	System	Mean (Dev)
Tilde	1.315 (0.041)	Tilde	1.265 (0.034)	Tilde	1.324 (0.033)
	1.275 (0.032)		*1.222 (0.026)*		*1.270 (0.034)*
S-CART	1.327 (0.036)	S-CART	1.294 (0.032)	S-CART	1.309 (0.044)
	1.265 (0.023)		*1.249 (0.018)*		*1.235 (0.042)*
M5′	1.191 (0.023)				
	1.177 (0.017)				
Lin.Reg.	1.411 (0.040)				
	1.390 (0.042)				

Global + P1 + P2 + R	
System	Mean (Dev)
Tilde	1.335 (0.036)
	1.283 (0.034)
S-CART	1.301 (0.049)
	1.253 (0.042)

positive) and TP reflects the true positive rate (proportion of positives predicted positive). The ideal case is FP = 0 and TP = 1. A classifier is represented by one (FP, TP) point in an ROC diagram. Points to the upper left

are strictly better; points to the upper right or lower left may be better or not depending on the costs assigned to each type of error.

Numerical predictors can be turned into classifiers by choosing a threshold (a prediction above this threshold counts as positive). By varying the threshold, the classifier can be tuned towards higher TP or lower FP. Thus a regression model typically gives rise to a curve in the ROC diagram.

In our ROC analysis the positive class is *degradable* (hence true positives are degradable instances predicted degradable; false positives are resistant instances predicted degradable).

Figure 2 compares ROC curves and points for all classification and regression systems for backgrounds Global and P1 + P2. Figure 3 compares ROC curves for relational learners, backgrounds R, and P1 + P2 + R. The curves shown were obtained for one single partitioning (the first one). Curves for other partitionings were similar, though not exactly the same (e.g., this curve suggests that S-CART performs slightly better on regression than Tilde, and slightly worse for classification, but this is not consistently the case for other partitionings).

- Observation 3: No large differences between regression and classification are noticeable in general, although the best regression systems do beat the best classification systems.
- Observation 4: At first sight, the ROC curves seem contradictory to the results in Tables 3 and 4. Note that the linear regression ROC curve is the best one on the $P1 + P2$ diagram (this was also the case for other partitionings), while according to the tables, linear regression seems to perform worse than the other systems. Even though it did not consistently perform significantly worse than other systems, on partitioning one (depicted in the ROC diagram), linear regression performed significantly worse than S-CART, according to our statistical tests. This is absolutely unsupported by the ROC diagram.

Discussion

With respect to comparisons between different systems and data representations, the results of our first batch of experiments are mainly negative: We have not been able to show a clear difference in performance between the different systems, or between the different backgrounds (except for the Global background, which clearly contains too little information to make good predictions possible).

The background R contains relational information, whereas $P1$ and $P2$ are propositionalizations of relational information. The lack of significant differences between backgrounds suggests that each of these backgrounds in itself provides a sufficiently complete description of a compound from

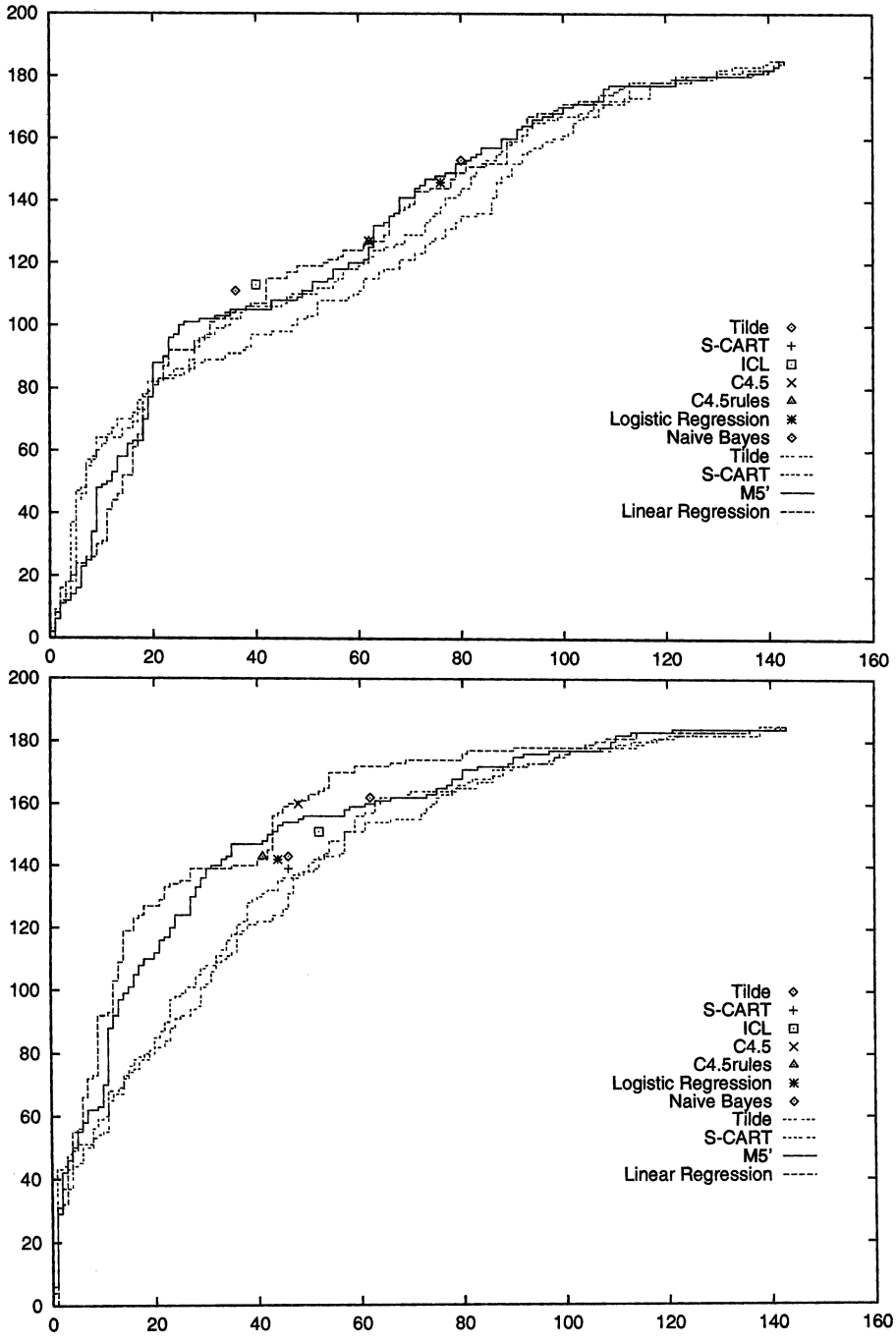

FIGURE 2 ROC curves comparing classification and regression systems: a) for background Global; b) for background $P1 + P2$. The horizontal axis represents the number of false positives, the vertical axis the number of true positives.

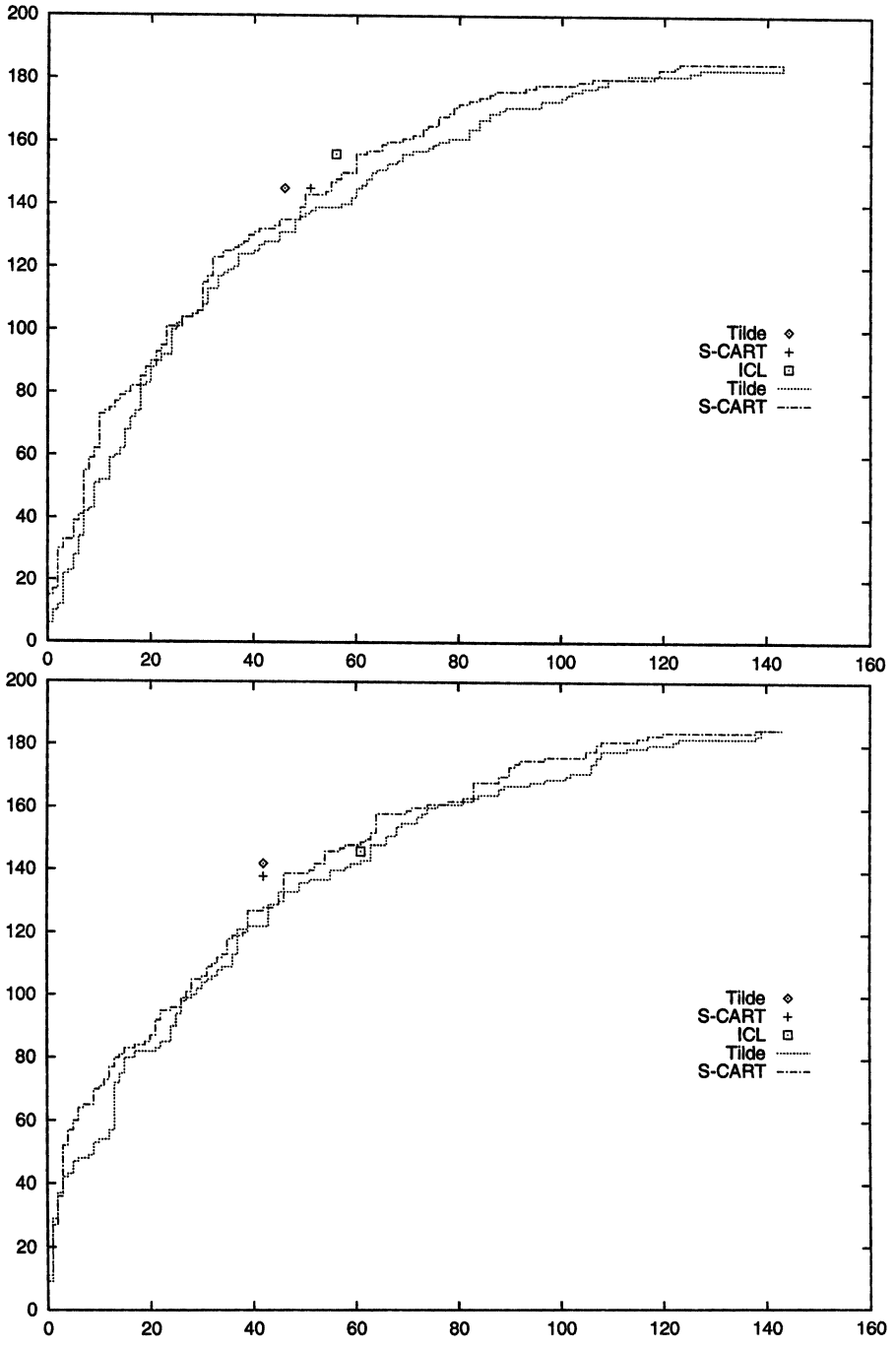

FIGURE 3 ROC curves comparing relational classification and regression systems: a) for background R; b) for background $P1 + P2 + R$. The horizontal axis represents the number of false positives, the vertical axis the number of true positives.

the viewpoint of predicting its degradability. Adding more information (by combining backgrounds) does not improve performance.

The difference between ROC curves and correlations or RMSEs is interesting. It can be explained from the observation that ROC curves essentially evaluate predictions from a classification point of view: If a hypothesis makes a large error in the numerical sense but this does not cause the instance to be misclassified, then the large error will decrease the correlation coefficient (increase the RMSE), but on the ROC curves this will not have a negative effect. One consequence of this is that outliers have a much more disturbing effect on the correlation/RMSE scores than on the ROC curves. For instance, in one case (for background $P1 + P2$), we found that removing three outliers reduced the RMSE of linear regression from 1.45 to 1.22, which immediately brought linear regression at the same level as M5′ (better than other systems).

Figure 4 compares predictions of linear regression and Tilde. The horizontal axis represents actual values, and the vertical axis represents predictions. It is clear from this figure what is happening: While both learners make predictions that are clearly correlated with the actual values, Tilde is in a sense more cautious. All its predictions are relatively close to the center. Linear regression predicts more extreme values, and extreme values have a large influence on both correlation and RMSE, which in this case causes linear regression to score very badly on these evaluation measures. For ROC curves, these extreme predictions do not hurt at all.

Our conclusion here is that one should be careful when choosing an evaluation criterion, and preferably not rely on a single one. The use of correlations and RMSEs to evaluate predictions is most appropriate when accurate numerical predictions are important, which may not always be the case. For example, in the biodegradability domain, a precise numerical prediction may not be that important, one is mainly interested in the area in which a compound lies (this is especially true in our case, where target values are

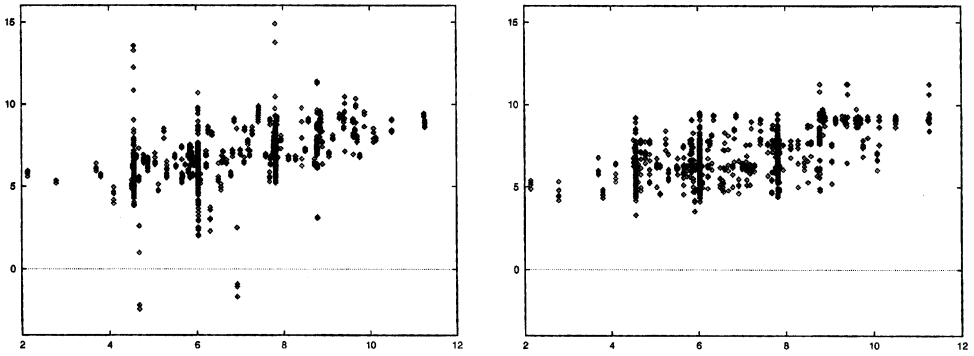

FIGURE 4 Predictions made by linear regression (left) and Tilde (right) in function of actual value, for the $P1 + P2$ background.

expert estimates and not measured values). Another point is that criteria may be very sensitive to outliers, as demonstrated here for correlation and RMSE.

Second Batch: Prediction of Intervals

After the first batch of experiments had been run, and noticing that combining information from different chunks of background knowledge does not improve performance, the question arised whether and how these results could be improved. Combining the hypotheses from different systems did not improve performance. A technique that did improve performance is the following one.

The HLTs, which formed our target values, are actually derived from upper and lower bound estimates by domain experts (we just took the arithmetic mean). An alternative way of predicting them is to build predictive models for these upper and lower bounds, and then build a predictive model for the mean that just predicts the mean of the upper and lower bound predictions.[3]

Except for the fact that with this approach more information is obtained (prediction of upper and lower bounds as well as the mean), the approach actually turns out to consistently yield equal or better performances, as can be seen in Tables 3 and 4, where the results of this batch of experiments is shown in italics. The results on ROC curves in general are also positive, though differences are quite small here.

That prediction of upper and lower bounds yields better results is not too surprising. Given that the original target value is derived from two other values, it seems reasonable to assume that there is a stronger relationship between these values and the structure of a compound, than between the mean of these two and the compound's structure.

Interpretation of Hypotheses

Next to predictive accuracy, comprehensibility of a hypothesis for domain experts is also important. We sent some of the produced hypotheses to our domain expert B. Kompare, who commented on them. Figure 5 shows a typical tree produced by Tilde together with some comments by the expert. An interpretation in plain English of the first leaf (labeled "very evident"), for instance, would be: "If the molecule's logP value is at least 4.84 and it contains a chlorine atom, then the molecule is resistant." The second leaf on which the expert commented can be described as containing molecules with a logP value between 1.67 and 4.84 that contain chlorine, benzene, and an R-Halogen where R is part of a resonant ring, but no non-resonant C_5 ring or methyl. There are 14 such molecules in the data set, of which 13 are resistant. The expert's comments suggest that of all these conditions, the *benzene* condition is the most relevant one.

```
activ_2(A,B)
logP(A,C), C >= 1.67 ?
+--yes:atm(A,D,cl,E,F) ?
|      +--yes:logP(A,G), G >= 4.84 ?
|      |      +--yes:[resistant] [16 / 16] ... very evident, OK, high logP ==> resistant
|      |      +--no: non_ar_5c_ring(A,I,J) ?
|      |             +--yes:[degrades] [2 / 2]
|      |             +--no: benzene(A,L,M) ?
|      |                    +--yes:ar_halide(A,N,O) ?
|      |                    |      +--yes:methyl(A,P,Q) ?
|      |                    |      |      +--yes:[degrades] [3 / 4]
|      |                    |      |      +--no: [resistant] [13 / 14] ... , Ok, benzenes are resistant
|      |                    |      +--no: [degrades] [4 / 4]
|      |                    +--no: [resistant] [28 / 31]
|      +--no: logP(A,V), V >= 4.84 ?
|             +--yes:ester(A,W,X) ?
|             |      +--yes:[degrades] [5 / 5]
|             |      +--no: atm(A,Z,n,A_1,B_1) ?
|             |             +--yes:amine(A,C_1,D_1) ?
|             |             |      +--yes:[resistant] [2 / 2]
|             |             |      +--no: [degrades] [2 / 2]
|             |             +--no: [resistant] [15 / 15] ... looks OK, i.e. other slowly degradable
|             +--no: mweight(A,H_1), H_1 >= 110.971 ?
|                    +--yes:n2n(A,I_1,J_1) ?
|                    |      +--yes:[degrades] [3 / 3]
|                    |      +--no: ester(A,L_1,M_1) ?
|                    |             +--yes:[degrades] [7 / 8]
|                    |             +--no: methoxy(A,O_1,P_1) ?
|                    |                    +--yes:[resistant] [3 / 3]
|                    |                    +--no: methyl(A,R_1,S_1) ?
|                    |                           +--yes:atm(A,T_1,n,U_1,V_1) ?
|                    |                           |      +--yes:atm(A,W_1,o,X_1,Y_1) ?
|                    |                           |      |      +--yes:sbond(A,T_1,Z_1,A_2), atm(A,Z_1,h,B_2,C_2) ?
|                    |                           |      |      |      +--yes:[degrades] [3 / 4]
|                    |                           |      |      |      +--no: [resistant] [7 / 9]
|                    |                           |      |      +--no: [resistant] [3 / 3]
|                    |                           |      +--no: [degrades] [6 / 8]
|                    |                           +--no: atm(A,F_2,o,G_2,H_2) ?  ... -O- is easy place to attack
|                    |                                  +--yes:[degrades] [13 / 15] ... logical
|                    |                                  +--no: five_ring(A,I_2,J_2) ?
|                    |                                         +--yes:[resistant] [3 / 3]
|                    |                                         +--no: six_ring(A,L_2,M_2) ?
|                    |                                                +--yes:[degrades] [9 / 12] ... ? cannot check
|                    |                                                +--no: [resistant] [2 / 2]
|                    +--no: [degrades] [12 / 13] ... Ok, pretty obvious 1.67<logP<4.84 and molW<111
+--no: atm(A,P_2,n,Q_2,R_2) ?
       +--yes:sbond(A,P_2,S_2,T_2), atm(A,S_2,o,U_2,V_2) ?  ?cannot check! is this -S=O group?
       |      +--yes:[resistant] [14 / 18]   ... identified exceptions - could not check!
       |      +--no: imine(A,W_2,X_2) ?
       |             +--yes:[resistant] [2 / 2]    ... identified exceptions
       |             +--no: methoxy(A,Z_2,A_3) ?
       |                    +--yes:[resistant] [2 / 2]  ... identified exceptions
       |                    +--no: [degrades] [30 / 33]  ...OK
       +--no: [degrades] [57 / 62] ... OK, pretty obvious, light organic of only C,H,(O), no N
```

FIGURE 5 Tilde classification tree, with comments added by expert.

The time needed by the expert to interpret the tree in Figure 5 was in the order of half an hour. It is clear from the expert's comments that some of the expert's knowledge is rediscovered by the tree and the expert can recognize this in the tree; the expert can even link some tests to specific chemical substructures. Also, the size of the tree is manageable.

Our conclusions from the expert's comments are that a) the trees and rules typically produced for this application are sufficiently interpretable, and b) the expert had a preference for rules over trees, mainly because of the possibility to interpret each rule separately; however, he added that a nicely structured single-page presentation of a tree helps a lot in interpreting it.

An interesting observation is also that ILP systems tend to produce relatively small models, e.g., regression trees produced by S-CART and Tilde typically contain around 50, respectively, 30 nodes, whereas trees produced by M5′ contain around 300 nodes on the average. It is not completely clear why this happens, as the smaller trees also occur for propositional data; it seems to be a property of the systems rather than the approach (trees for a relational background still tend to be slightly smaller than for a propositional one, but this difference is much smaller). Obviously smaller theories are preferred by domain experts, if this can be achieved without loss of predictive accuracy this is an advantage.

Comparison with the BIODEG Program

Howard and Meylan (1992) describe the BIODEG program for biodegradability prediction. This program estimates the probability of rapid aerobic biodegradation in the presence of mixed populations of environmental organisms. It uses a model derived by linear regression (Howard et al. 1992). Here we compare our results with this program. It should be noted that some of the compounds in our data set were used to train the BIODEG program, so there is an overlap of training and test set, which puts BIODEG at an advantage.

On our data set, the predictions of the BIODEG program have a correlation of 0.607 with the actual values. Most of the correlations we have obtained are considerably higher. In the light of the "correlation vs. ROC" discussion, this result should be interpreted with caution. Indeed, the BIODEG ROC curve is better than the curves we obtain in most of our experiments; however, it is worse than our own linear regression method on the $P1 + P2$ background, as Figure 6 shows.[4]

The conclusion here is that whether accurate numerical predictions are important or not (in other words, whether correlations and RMSEs are the main evaluation criterion, or whether ROC curves are), in both cases we have an approach that outperforms the BIODEG program.

RELATED WORK

The observation that propositional learners with a good propositional description of compounds may perform as well as ILP is consistent with Srinivasan and King's work (Srinivasan and King 1997) on the use of ILP-induced features in linear regression. The difference between their approach and ours is that Srinivasan and King generate propositional features in a less trivial way than we do; they use an ILP system to generate features deemed relevant by the system, whereas we have used less sophisticated or more human-controlled feature generators (e.g., just counting all substructures of a certain kind, or counting all substructures considered relevant by chemists).

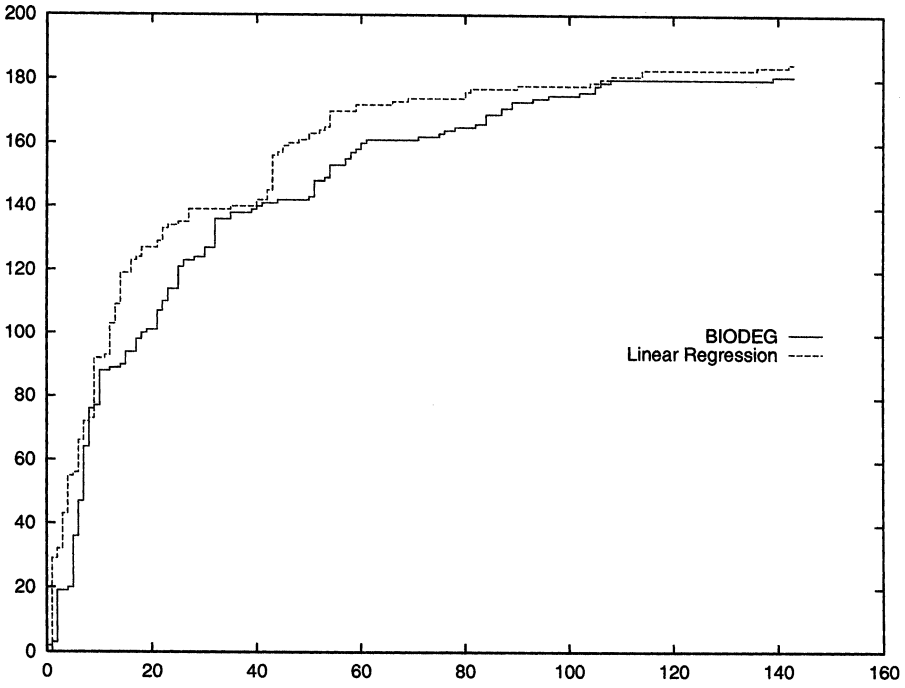

FIGURE 6 ROC curve for BIODEG program, compared to ROC curve of linear regression on $P1 + P2$. The horizontal axis represents the number of false positives, the vertical axis the number of true positives.

Our results suggest that, from the machine learning point of view, even such trivial propositionalizations may work well.

Other related work includes QSAR applications of machine learning and ILP, on one hand, and constructing QSAR models for biodegradability, on the other hand. On the ILP side, QSAR applications include drug design (King et al. 1992), mutagenicity prediction (Srinivasan et al. 1996), and toxicity prediction (Srinivasan et al. 1997). The latter two are closely related to our application. In fact, we have used a similar representation and reused parts of the background knowledge developed for them.

On the biodegradability side, the work by Howard et al. (1992) is closest to our work. The BIODEG program for biodegradability prediction (Howard and Meylan 1992) estimates the probability of rapid aerobic biodegradation in the presence of mixed populations of environmental organisms. It uses a model derived by linear regression (Howard et al. 1992). The results presented in this paper show that it is possible to improve upon these results, whether correlations or ROC curves are used as the evaluation measure.

Work on applying machine learning to predict biodegradability includes a comparison by Kompare (1995) of several AI tools on the same domain and data; he found these to yield better results than the classical statistical and

probabilistic approaches. Zitko (1991) and Cambon and Devillers (1993) applied neural nets, and Gamberger et al. (1987) applied several approaches.

CONCLUSIONS

This paper presents a case study on the use of machine learning algorithms for the prediction of biodegradability of chemical compounds. We have performed experiments with a wide range of algorithms, for a variety of background information, using different approaches. Our main conclusions are:

- When evaluating systems based on their predictive accuracy, correlation, or RMSE, it does not seem to matter very much which system is used; all of them perform very similarly. Linear regression may seem to lag somewhat behind, but this can be attributed to the sensitivity of RMSE and correlation to outliers.
- Because regression trees make more cautions predictions (closer to the global mean) than some other methods such as regression, a comparison based on correlations or RMSEs puts them at an advantage. *Correlations should be interpreted with caution when used to compare different regression approaches*.
- ROC curves may give a very different impression than correlations or RMSEs. In our case, ROC curves suggest that linear regression based on automatically generated features works very well.
- Propositionalization of structural descriptions is a good alternative to the direct use of relational background knowledge. It has the advantage that more predictive modeling techniques are available for propositional knowledge than for relational knowledge, cf. Srinivasan and King (1996).
- Indirect ways of building predictive models may be useful in obtaining better performance. In our case, separate prediction of lower and upper bounds (from which the mean is computed afterwards) turns out to yield slightly better models than direct prediction of HLTs. The improvement is especially noticeable when using a relational background.

The prediction of intervals is in itself an interesting research topic in machine learning, even besides the fact that it may yield better predictions of a mean value; in some application domains experts are more interested in intervals than in point predictions. As such, this seems an interesting topic for future research.

To conclude, we believe that this case study has pointed out some interesting issues concerning the use of machine learning and statistical methods for QSAR modeling, and also some more general issues concerning the relationship between classification and regression approaches and ways to evaluate them.

NOTES

1. An alternate approach would have been to use a more formal common declarative bias language; recent developments in this direction are described by Knobbe et al. (2000).
2. Given the large number of comparisons, some significant results are expected; we say that a difference is consistently significant if it is significant for all five partitionings for which a cross-validation was performed.
3. More precisely, predictions on a logarithmic scale are first transformed back to the original scale, then the mean is computed, then the logarithm of this is taken.
4. Due to a few missing predictions for BIODEG, the curves do not end in the same point; however, in the best case for BIODEG, when all missing predictions would have been correct, its whole curve just shifts a bit upwards but not enough to change the outcome of this comparison.

REFERENCES

Blockeel, H., and L. De Raedt. 1998. Top-down induction of first order logical decision trees. *Artificial Intelligence* 101(1–2):285–297.

Breiman, L., J. Friedman, R. Olshen, and C. Stone. 1984. *Classification and Regression Trees*. Belmont: Wadsworth.

Cambon, B., and J. Devillers. 1993. New trends in structure-biodegradability relationships. *Quant. Struct. Act. Relat*. 12(1):49–58.

Clark, P., and R. Boswell. 1991. Rule induction with CN2: Some recent improvements. In *Proceedings of the Fifth European Working Session on Learning*, ed. Y. Kodratoff, Volume 482 of *Lecture Notes in Artificial Intelligence*, pages 151–163. Springer-Verlag.

De Raedt, L., and W. Van Laer. 1995. Inductive constraint logic. In *Proceedings of the Sixth International Workshop on Algorithmic Learning Theory*, eds. K. P. Jantke, T. Shinohara, and T. Zeugmann, Volume 997 of *Lecture Notes in Artificial Intelligence*, pages 80–94. Springer-Verlag.

Dehaspe, L., and H. Toivonen. 1999. Discovery of frequent datalog patterns. *Data Mining and Knowledge Discovery* 3(1):7–36.

Dietterich, T. G. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation* 10(7):1895–1924.

Džeroski, S., H. Blockeel, S. Kramer, B. Kompare, B. Pfahringer, and W. Van Laer. 1999. Experiments in predicting biodegradability. In *Proceedings of the Ninth International Workshop on Inductive Logic Programming*, eds. S. Džeroski and P. Flach, Volume 1634 of *Lecture Notes in Artificial Intelligence*, pages 80–91. Springer-Verlag.

Gamberger, D., S. Sekuak, and A. Sabljič. 1987. Modelling biodegradation by an example-based learning system. *Informatica* 17:157–166.

Howard, P., R. Boethling, W. Jarvis, W. Meylan, and E. Michalenko. 1991. *Handbook of Environmental Degradation Rates*. Chelsea, MI: Lewis Publishers.

Howard, P., R. Boethling, W. Stiteler, W. Meylan, A. Hueber, J. Beauman, and M. Larosche. 1992. Predictive model for aerobic biodegradability developed from a file of evaluated biodegradation data. *Environ. Toxicol. Chem*. 11:593–603.

Howard, P., and W. Meylan. 1992. User's guide for the biodegradation probability program, ver. 3. Technical report, Syracuse Res. Corp., Chemical Hazard Assessment Division, Environmental Chemistry Center, Syracuse, NY 13210, USA.

King, R., S. Muggleton, R. Lewis, and M. Sternberg. 1992. Drug design by machine learning: The use of inductive logic programming to model the structure-activity relationships of trimethoprim analogues

binding to dihydrofolate reductase. In *Proceedings of the National Academy of Sciences* 89(23), pages 11322–11326, National Academy of Sciences, Washington, DC.

Knobbe, A., A. Siebes, H. Blockeel, and D. van der Wallen. 2000. Multirelational data mining, using UML for ILP. In *Proceedings of The Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases* (*PKDD 2000*), Volume 1910 of *Lecture Notes in Artificial Intelligence*, pages 1–12, Lyon, France. Springer.

Kompare, B. 1995. *The Use of Artificial Intelligence in Ecological Modelling*. Ph. D. thesis, Royal Danish School of Pharmacy, Copenhagen, Denmark.

Kramer, S. 1996. Structural regression trees. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pages 812–819, Cambridge/Menlo Park. AAAI Press/The MIT Press.

Kramer, S. 1999. *Relational Learning vs. Propositionalization: Investigations in Inductive Logic Programming and Propositional Machine Learning*. Ph. D. thesis, Vienna University of Technology, Vienna, Austria.

Provost, F., and T. Fawcett. 1998. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, pages 43–48. AAAI Press.

Quinlan, J. 1993a. Combining instance-based and model-based learning. In *Proceedings of the 10th International Workshop on Machine Learning*, pages 236–243. San Francisco, CA: Morgan Kaufmann.

Quinlan, J. R. 1993b. *C4.5: Programs for Machine Learning*. Morgan Kaufmann series in Machine Learning. San Francisco, CA: Morgan Kaufmann.

Srinivasan, A., and R. King. 1997. Feature construction with inductive logic programming: A study of quantitative predictions of biological activity by structural attributes. In *Proceedings of the Sixth International Workshop on Inductive Logic Programming*, Volume 1314 of *Lecture Notes in Artificial Intelligence*, pages 89–104. Springer-Verlag.

Srinivasan, A., R. King, S. Muggleton, and M. Sternberg. 1997. Carcinogenesis predictions using ILP. In *Proceedings of the Seventh International Workshop on Inductive Logic Programming*, Lecture Notes in Artificial Intelligence, pages 273–287. Springer-Verlag.

Srinivasan, A., S. Muggleton, M. Sternberg, and R. King. 1996. Theories for mutagenicity: A study in first-order and feature-based induction. *Artificial Intelligence* 85(1,2):277–299.

Wang, Y., and I. Witten. 1997. Inducing model trees for continuous classes. In *Proceedings of the 9th European Conf. on Machine Learning Poster Papers*, pages 128–137, Prague, Czech Republic.

Weininger, D. 1988. SMILES, a chemical and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* 28(1):31–36.

Zitko, V. 1991. Prediction of biodegradability of organic chemicals by an artificial neural network. *Chemosphere* 23(3):305–312.